Speech Prosody 2024
Book of abstracts

RMT: Tools for Prosodic Computational Literacy

Dafydd Gibbon

Prosodic computational literacy is an important goal for students of acoustic phonetics, especially those from endangered language communities in less affluent countries. There are several convenient 'off-the-shelf' packages for prosody computation, including Praat, ProsodyPro, Prosogram, ProZed, Winpitch, and many convenient Praat scripts. However, experiments typically require small hybrid intersections of functionalities of these packages together with spreadsheets, R, Praat scripting or Python. Python was chosen in order to enable non-hybrid, seamless embedding of small tools into larger systems for exploratory research, because of scalability, and because of the availability of extensive Python libraries to support in-depth in-sight into filters and transformations rather than using ready-made complex functionalities. A design criterion for the toolkit is overall coherence and clarity of structure. The tools cover the analysis of speech signal annotations, and a modulation-theoretic approach to the demodulation of speech signal amplitude modulation and frequency modulation. Comparison of results is enabled by provision of distance measurement and hierarchical clustering techniques. The approach has been evaluated in practice in a range of publications and in teaching.

How pupillary responses reflect the predictability of word stress in Turkish

İpek Pınar Uzun

Processing of word stress is a complex mechanism that has an impact on prosodic information, the predictability of speakers' native language, and their sensitivity to shifting stress positions. It is still unclear whether the complexity arises in response to alternations in Turkish stress patterns and whether they exhibit sensitivity to syllable structure. To address these questions, two studies were designed to focus on the pupil dilation response (PDR), a measure employed as a proxy for cognitive load. The current study examined the changes in pupil size with forty-six native Turkish speakers. In Experiment 1, 30 CV.CV.CV words were used with canonical final stress and 30 words in a 2×3 design with penultimate stress pronounced with final, medial, and initial stress types each. In Experiment 2, a syllable weight factor was added by using 30 CVC.CV.CV words in a 3×3 design from each of the three stress types with shifts to other syllables. The LME models showed that the changes from exceptional stress to a default final position do not directly cause increased PDR for Experiment 1. However, the results for Experiment 2 suggest that syllable weight is sensitive to PDR and plays a significant role in cognitive processing.

Interested-Sounding Voices Influence Listeners' Willingness to Disclose

Siti Aisyah Binti Shamshun Baharin, Veronica Lamarche, Netta Weinstein and Silke Paulmann

“Social prosody” refers to prosodic variations in speech that convey a specific social intention. We employed social prosody variations to study the impact of how the voice of “showing an interest” can have on others. Specifically, this research explores the nuanced influence of prosody on how listeners self-disclose and their communicative intentions (e.g., willingness to disclose). Across three studies (Study 1a native English Speakers; Study 1b native Asian speakers; Study 1c English & Asian speakers), we found that listeners reported feeling closer to speakers and demonstrated increased satisfaction and willingness to engage with speakers when they spoke in an “interested-sounding” voice compared to an "uninterested" one. This paper describes the third study, a cross-cultural comparison between British and Indian speakers.

Disclosure analysis shows listeners used higher word frequency (Word Count); higher logical thinking (Analytical Thinking); communicated with more perceived honesty (Authentic Communication); higher linguistic style and language comprehension (Linguistic Style), more words signalling cognitive engagement and use of formal language (Function Words), and pronouns reflecting social hierarchy and attentional focus dimension (Pronouns used) to speakers with interesting-sounding voices. These findings suggest that the way we show an interest in someone is key in influencing listeners’ emotional responses and communicative intentions.

Fundamental frequency in French-speaking children: about the influence of gender and social class

Erwan Pépiot

This study deals with the productions of 8- to 10-year-old French-speaking children living in Paris area in a reading task and in semi-spontaneous speech. Two groups of speakers were recorded: pupils from an upper-class private school (10 girls / 8 boys), and children studying in a lower-class state school (8 girls / 10 boys). Mean fundamental frequency and F0 modulation were measured. Results show that girls from both schools presented a significantly higher average F0 than boys in both tasks. This difference was slightly more pronounced among children from the privileged school. No significant correlation was found between the speakers' height and their average F0, which rules out any physiological explanation. F0 modulation was significantly higher for girls in the privileged school but was very similar between both genders in the disadvantaged school. Regardless of gender, F0 modulation was stronger in children from privileged backgrounds, which suggests that a large modulation of pitch in French could be an upper-class marker. Overall, these data support the idea that pre-pubescent children would tend to develop gender-related vocal practices by adapting their pitch and their intonation in order to match the differences observed in adults.

Intonation and Fluency in Emotionally Dysregulated French Patients with an Acquired Brain Injury: Case Studies

Thalassio Briand, Camille Fauth and Marie Kuppelin

Acquired brain injuries (ABI) often result in persistent emotional dysregulation. As part of a larger study, this paper explores the effectiveness of a dialectical behaviour therapy (DBT) programme, a type of evidence-based psychotherapy that helps patients build emotional regulation skills. Current literature suggests that a significant proportion of disfluency, faster speech rate and large amplitudes in frequency variation may denote high emotional activation in speech. However, identifying emotional regulation remains a challenge, as dysregulation may manifest as emotional apathy in addition to extreme activation. This complicates establishing a direct link between intonation patterns and emotional dysregulation. Two ABI patients were recorded in 40-90 minutes semi-directive interviews, in which they narrated emotionally charged memories and described pictures. One patient was recorded five months prior to and immediately before beginning the therapy programme (t0-t1), and the other before and after the five-month programme (t1-t2). Our findings highlight more variation in intonation at t1 for both patients, and the patient who followed the programme decreased his speech rates over time. These may be indicators of the therapy's effectiveness. Moreover, our results suggest that the ratio of disfluencies may not be a good indicator of emotional dysregulation in ABI patients.

Exploring natural speech intonation of an under-researched Papuan language

Alexander Zahrer

New Guinea is home to >800 languages. The poor general documentation of these languages is even worse when it comes to prosody. Existing accounts use read speech and staged conversations instead of natural speech. Field workers often lack time and experience for thorough prosodic analyses.

Contour clustering (Kaland 2021) is an effective method to approach under-researched languages. Our target language is Muyu (Zahrer 2019, 2023) for which we formulate the following research question: Does the syntactic type of verb correlate with f0 movements in the contour? There are three types of verbs: 'final verbs' indicate end of a sentence, whereas sentences continue after 'medial verbs' and 'multi verbs'. To test a possible correlation, we segmented contours from spontaneous speech and annotated them for the contained verb type.

The results show a clear correlation. Hierarchical clustering of f0 time-series assigned a total of 319 contours to seven clusters. Mean contours were calculated for each cluster to indicate the typical pitch movement. Final verbs are assigned to falling pitch clusters, medial/multi verbs tend to rising or level pitch ($X^2(4, N=319)=38.114, p<.00001$). Therefore, Muyu uses f0 movement along with morphosyntactic marking of sentence (in)completeness.

Prosodic realization of different focus types in Persian

Simon Roessig, Mortaza Taheri-Ardali, Lena Pagel and Doris Mücke

This paper presents results from a production study on the marking of different focus types in Persian: broad, narrow, and corrective focus. Prior research has primarily concentrated on the distinction between broad and corrective focus. In the present work, we aim to disentangle the focus breadth and correctiveness. In addition, we take a closer look at all syllables of the focused target word. Based on a data set of 12 native speakers of Persian involved in an interactive game-like task, our analyses show the following: (1) Albeit being subtle, a differentiation between broad, narrow, and corrective focus is present regarding syllable duration; (2) F0-related parameters and intensity are only reliably modulated between broad focus on the one hand and narrow and corrective focus on the other hand; (3) both syllables of the target word are affected but with different patterns of cue modulations.

The interplay between syllabic duration and melody to signal prosodic functions in reading and story retelling in Brazilian Portuguese

Plinio Barbosa

This research examined the relationship between Thai prosodic words and musical note pairs, focusing on prominence matching and quantity matching. We hypothesized that the iambic structure of Thai disyllables would result in matching with more prominent musical positions and longer note duration. An analysis of 40 most frequently used disyllables from a 4,078,300 word corpus of Thai pop lyrics however revealed that prominence is not the main factor determining Thai textsetting. The sampled disyllables were found to match with initial prominence and final prominence note pairs roughly equally (48.8% v. 51.1%). On the contrary, quantity matching played a crucial role, with disyllables preferring to map to short- long note pairs (72.8%) especially at the end of musical phrases (92.2% of all phrase-final cases). Nevertheless, prominence matching plays a secondary role, resulting in the tendency of even disyllables to align with the prominence-final note pairs (76.8% of all even pair). This study thus demonstrates that, in contrast to previously studied languages, final stressed syllables in Thai prosodic words match with extended note quantity in textsetting, making musical prominence secondary. The importance of note quantity over musical prominence suggests the Iambic/Trochaic Law's role in the language-music interaction.

Prosodic grouping in Akan and the applicability of the iambic-trochaic law

Constantijn Kaland, Anjali Bhatara, Natalie Boll-Avetisyan and Thierry Nazzi

The current study investigates prosodic grouping in Akan (iso: aka), a Kwa language. Except for tone, the prosody of Akan is largely understudied. Even fewer studies have addressed the perception of prosody. The current study investigates how prosodic cues such as duration and intensity contribute to grouping in this language by a replication of a perception experiment used in earlier work. In this experiment, participants indicate whether auditory sequences are composed of repetitions of weak-strong or strong-weak sound pairs. The experiment was designed to test the predictions of the iambic-trochaic law (ITL). The ITL predicts that auditory sequences alternating in duration are grouped as iambs (weak-strong), whereas sequences alternating in intensity are grouped as trochees (strong-weak). In addition to the role of the acoustic cues of duration and intensity, the current study also tests how overall acoustic variability in the sequences affects prosodic grouping. The results for Akan show that duration differences lead to iambic grouping, as predicted by the ITL, but only when there is low acoustic variability in the sequences. No grouping effects for intensity were found. The results are interpreted and discussed in a typological context.

Prosodic Word Stress in Text-Setting in Thai Pop Songs

Komtham Domrongchareon and Pittayawat Pittayaporn

This research examined the relationship between Thai prosodic words and musical note pairs, focusing on beat falls and note duration. We hypothesized that the iambic structure of Thai disyllables would result in alignment with more prominent musical positions and longer note durations. An analysis of 40 most frequently used disyllables from a 4,078,300 word corpus of Thai pop lyrics however revealed that prominence is not the main factor determining Thai text setting. The sampled disyllables were found to align with prominence-initial and prominence-final note pairs roughly equally (51.1% v. 48.8%). On the contrary, duration alignment played a crucial role, with disyllables preferring to map to short-long note pairs (72.8%), especially at the end of musical phrases (92.2% of all phrase-final cases). Nevertheless, prominence alignment plays a secondary role, resulting in the tendency of even disyllables to align with the prominence-final note pairs (76.8% of all even pair). This study thus demonstrates that, in contrast to previously studied languages, final stressed syllables in Thai prosodic words align with extended note durations in text setting, making musical prominence secondary. The importance of note duration over musical prominence suggests the Iambic-Trochaic Law's role in the language-music interaction.

How to test gesture-speech integration in ten minutes

Matteo Maran and Hans Rutger Bosker

Human conversations are inherently multimodal, including auditory speech, visual articulatory cues, and hand gestures. Recent studies demonstrated that the timing of a simple up-and-down hand movement, known as a beat gesture, can affect speech perception. A beat gesture falling on the first syllable of a disyllabic word induces a bias to perceive a strong-weak stress pattern (i.e., “CONtent”), while a beat gesture falling on the second syllable combined with the same acoustics biases towards a weak-strong stress pattern (“conTENT”). This effect, termed the “manual McGurk effect”, has been studied in both in-lab and online studies, employing standard experimental sessions lasting approximately forty minutes. The present work tests whether the manual McGurk effect can be observed in an online short version (“mini-test”) of the original paradigm, lasting only ten minutes. Additionally, we employ two different response modalities, namely a two-alternative forced choice and a visual analog scale. A significant manual McGurk effect was observed with both response modalities. Overall, the present study demonstrates the feasibility of employing a ten-minute manual McGurk mini-test to obtain a measure of gesture-speech integration. As such, it may lend itself for inclusion in large-scale test batteries that aim to quantify individual variation in language processing.

Tone Value Representation for Computer-Assisted Pronunciation Training

Wu-Hao Li, Te-Hsin Liu and Chen-Yu Chiang

Scholars have long demonstrated the effectiveness of using visual representations to aid non-native speakers in learning Mandarin tone. Among these methods, presenting pitch contours visually has proven to be the most intuitive and efficient. However, the variability of pitch contours in continuous speech, influenced by coarticulation, poses challenges for learners to follow accurately. To address this issue, we propose a method to normalize pitch contours, mitigating the effects of coarticulation and prosodic phrasing. Experimental results show that subjects can identify the class of tone by looking at the representation proposed in this study and evaluating the quality of the tones of the syllables pronounced by the speakers visually. In addition, our approach can present the mispronunciation with the contours of the relative pitch of the syllables. We believe that our method offers more comprehensive feedback to learners compared to relying solely on a tone recognition model for pronunciation evaluation.

The Perception of Declarative and Interrogative Sentences of Chinese Autistic Children

Yao Lu, Ruiyao Zhong and Xiyu Wu

This paper analyzes perceptual disparities between children with autism spectrum disorders (ASD) and their typically developing (TD) peers with respect to Chinese interrogative and declarative sentences. By synthesizing a continuum from interrogative to declarative sentences and conducting an identification experiment, this research systematically analyzes these differences. The experimental results reveals relatively significant differences in performance between both groups. These findings indicate that when teaching children with ASD to answer questions, educators need to employ more emphatic prosody in their questioning to achieve the intended interrogative effect.

Knowledge of a talker’s f0 affects subsequent perception of voiceless fricatives

Orhun Ulusahin, Hans Rutger Bosker, James M. McQueen and Antje S. Meyer

The human brain deals with the infinite variability of speech through multiple mechanisms. Some of them rely solely on information in the speech input (i.e., signal-driven) whereas some rely on linguistic or real-world knowledge (i.e., knowledge-driven). Many signal-driven perceptual processes rely on the enhancement of acoustic differences between incoming speech sounds, producing contrastive adjustments. For instance, when an ambiguous voiceless fricative is preceded by a high fundamental frequency (f0) sentence, the fricative is perceived as having lower a spectral center of gravity (CoG). However, it is not clear whether knowledge of a talker’s typical f0 can lead to similar contrastive effects. This study investigated a possible talker f0 effect on fricative CoG perception. In the exposure phase, two groups of participants (N=16 each) heard the same talker at high or low f0 for 20 minutes. Later, in the test phase, participants rated fixed-f0 /?ɔk/ tokens as being /sɔk/ (i.e., high CoG) or /ʃɔk/ (i.e., low CoG), where /?/ represents a fricative from a 5-step /s/-/ʃ/ continuum. Surprisingly, the data revealed the opposite of our contrastive hypothesis, whereby hearing high f0 instead biased perception towards high CoG. Thus, we demonstrated that talker f0 information affects fricative CoG perception.

Interpretation of Spanish stress by second language learners

Izaro Bedialauneta Txurruka

Stress is contrastive in both English and Spanish. Spanish uses intensity, pitch, and duration as acoustic correlates of stress, the strongest cues being F0 and duration in combination. In English, vowel reduction is the main cue of stress. Due to these differences L1 English/L2 Spanish speakers show difficulties perceiving stress in Spanish. Our goal is to compare the interpretation of Spanish stress by L2 to native speakers of Spanish. This study investigates the influence on the correct identification of the stress pattern within the word, as well as the impact of the location of the target word within the sentence. The study involved a 48-stimuli Forced-choice task in Qualtrics. 95 individuals participated: 29 L1 speakers, 14 L2 advanced, 21 L2 intermediates, and 31 L2 beginners. Descriptive statistics and a regression model suggest that as L2 speakers advance in their proficiency in Spanish, their ability to interpret stress improves. Moreover, until they reach an intermediate level of language proficiency, L2 learners are not able to extract the correct meaning of words minimally differing in stress pattern.

The Power of Prosody and Prosody of Power: An Acoustic Analysis of Finnish Parliamentary Speech

Martti Vainio, Antti Suni, Juraj Šimko and Sofoklis Kakouros

Parliamentary recordings provide a rich source of data for studying how politicians use speech to convey their messages and influence their audience. This provides a unique context for studying how politicians use speech, especially prosody, to achieve their goals. Here we analyzed a corpus of parliamentary speeches in the Finnish parliament between the years 2008-2020 and highlight methodological considerations related to the robustness of signal based features with respect to varying recording conditions and corpus design. We also present results of long term changes pertaining to speakers' status with respect to their party being in government or in opposition. Looking at large scale averages of fundamental frequency - a robust prosodic feature - we found systematic changes in speech prosody with respect opposition status and the election term. Reflecting a different level of urgency, members of the parliament have higher F0 at the beginning of the term or when they are in opposition.

Speech rate correlates with politeness in Spanish offers

Bruno Staszkiewicz

The current study investigates whether the variables of power, distance, and imposition can correlate with politeness by examining speech rate in Spanish offers. The hypothesis is that slower speech rates occur in more face-threatening situations. Participants, 34 native Spanish speakers, completed a contextualized sentence-reading task where they read aloud 8 paragraph-length contextualizing situations followed by an offer and repeated it three times across three blocks. The situations were balanced for two levels of power (high/low), distance (high/low), and imposition (high/low). The analysis of 762 offers focused on the syllable duration of the target sentences.

A linear mixed-effect model analysis was conducted in R to observe the effect of the contextual variables on the use of speech rate. Results showed that distance and imposition significantly influenced syllable duration, while power did not. The overall results indicate that speakers produced a slower speech rate when the interlocutors did not know each other as well as when the speakers offered to do something that was of high cost to accomplish for themselves.

Tonal density characterises the scope of the overbid use of connector “mais” in French conversation

Cristel Portes, Marie Kolenberg, Stéphane Rauzy and Roxane Bertrand

Most uses of mais (‘but’) as a connector in French (i.e. correction, opposition, concession) imply a contrast between their antecedent p and their scope q but differ in terms of argumentative orientation. Conversely, p and q in overbid mais (‘des grands mecs mais très grands!’ 'tall guys mais very tall!') present a co-oriented argumentation and the contrast is expressed by a semantic strengthening from p to q. We hypothesized that this should be reflected by both a lexical and a prosodic strengthening of the scope constituent q. To test this hypothesis, we compared the overbid use with the concessive use in conversational data. We found more adverbial intensifiers in the scope of the overbid than the concessive use. We also investigated both the phonetic and phonological realisation of mais itself and of its scope. We found that both mais are phrased with their scope rather than alone. Crucially, we found a significantly greater H tone rate (both per second and per syllable) on the scope of the overbid. We also found a lower articulation rate but only in the absence of adverbial intensifiers. This demonstrates the role of tonal density in prosodic strengthening.

Prosody of corrective "but" sentences in English

Danfeng Wu

This paper provides prosodic evidence for a syntactic analysis of corrective “but” sentences, and argues that prosodic structure is not completely flat, but can replicate the dominance relations in the syntax. Corrective “but” sentences are “but”-coordination that requires negation in the first conjunct, such as (1) “Max misses not spinach but chard” and (2) “Max doesn’t miss spinach but chard”.

There is debate about the syntactic analysis of (1): Toosarvandani (2013) analyzed it as DP-coordination (Max misses [not spinach] but [chard]), while Wu (2022) argued that it is structurally ambiguous between DP-, vP- and TP-coordination. In a production study, we showed with duration-based evidence that the prosodic boundary following the first conjunct (i.e. following “spinach”) is stronger than the boundary following a typical DP, supporting Wu’s analysis.

(2) is uncontroversially analyzed as vP-coordination plus ellipsis in the literature, and this vP embeds a DP (Max does not [miss [spinach]] but chard). In the second part of the study, we took advantage of this recursive syntactic structure in (2) and asked whether it might lead to recursive prosodic structure. Duration-based evidence suggests that it does, as the prosodic boundary following “spinach” is stronger than the boundary following an unembedded DP.

Variation in speech rhythm in Tongan English

Danielle Tod

The current paper investigates speech rhythm in the L2 variety of English spoken in the Kingdom of Tonga, an island nation in the South Pacific, using the quantitative metrics nPVI-V, rPVI-C and %V. In addition to classifying Tongan English in relation other varieties of English, external constraints on variation in speech rhythm are examined. A corpus of conversational and read speech recorded with 48 Tongan English speakers is used in analysis. Results indicate that inter-speaker variation exists, with education, occupation and index of English use are significant constraints on variation. The vocalic metrics, nPVI-V and %V, had greater explanatory power, proving to be more robust measures of variation than rPVI-C.

Perceptual cues to checked tones in Shanghai Chinese

Chengjia Ye and Paul Boersma

Shanghai Chinese has a complicated tone system conditioned by phonological onset voicing and syllable structure. Checked tones in Shanghai Chinese are acoustically realized with tenser phonation, shorter duration and a more central vowel space than their unchecked counterparts. The present study examines whether these cues also play a role in speech perception. A three-alternative forced-choice experiment was conducted in which 40 native listeners chose among three Chinese characters representing an open /CV/, a nasal-ending /CVN/ and a glottal-ending /CVʔ/ syllable according to the word they heard. Results show main effects of phonation type, vowel quality and duration as well as the interaction between the former two, on both decisions and reaction time. For reaction time, two-way interactions between duration and the other two cues were also detected. From the non-overlapping confidence intervals of the three main effects we tentatively conclude that tenser phonation is the primary perceptual cue to checked tones, while a more central vowel space and shorter duration are secondary, in line with findings in other tonal languages like Vietnamese and Cantonese. Additionally, we found that the low vowel pair /ɐ/ and /ɑ/ and high register tones tend to bias perception towards checked tones.

Your Prosody Matters! The Effect of Controlling Tone of Voice on Listeners across the lifespan

Berdien Vrijders, Netta Weinstein, Silke Paulmann, Bart Soenens, Joachim Waterschoot and Maarten Vansteenkiste

Following Self-Determination Theory, speakers can communicate either in more controlling or more autonomy-supportive ways. Whereas most previous studies focused on the content of both communication styles, the current research examined whether experimentally induced controlling versus autonomy-supportive tone of voice differentially predicts listeners' experienced autonomy frustration, closeness, intention for collaboration, and feelings of sadness and anger, even when listeners are exposed to these communications only briefly. In three studies with adults (Study 1, N = 61; Mage = 31.51), parents (Study 2, N = 111; Mage = 44.73), and toddlers (Study 3, N = 189; Mage = 4.93), multilevel analyses indicated that sentences spoken with a harsher (i.e., with increased vocal energy), relative to a softer tone of voice were perceived as more pressuring, leading to higher levels of experienced autonomy frustration (Study 1, 2 and 3). Listening to such harsh voices explained why listeners felt less close to and were less inclined to collaborate with controlling speakers (Study 2) and reported higher levels of anger and fear (Study 3). Results for the first time show the impact of speakers' tone of voice on listeners across ages, with adults and toddlers alike reporting more maladaptive effects following controlling tone of voice.

Mandarin-speaking 6-year-olds can use preboundary pitch range expansion to disambiguate compounds from lists

Feng Xu, Ping Tang, Katherine Demuth and Nan Xu Rattanasone

Pitch can be used for marking boundaries and chunking utterances into different units, e.g., compounds (N1+N2, e.g., jellybeans) and their corresponding list forms (N1, N2, e.g., jelly, beans). Unlike English, where 5-6-year-olds can use different pitch patterns to mark boundaries in an adult-like manner, Mandarin is a tonal language using pitch for both word meanings (lexical tone) and utterance meanings (e.g., preboundary pitch range expansion). While lexical tones are early acquired, it was unclear when Mandarin-speaking children can use pitch cues to disambiguate compounds and lists. A total of 41 adults and 29 6-year-olds participated in an elicited production task. The pitch range of N1 was measured (highest minus lowest pitch (f0)). Our results showed that similar to adults, 6-year-olds produced a larger pitch range over N1 in lists compared to compounds for rising and falling tones. However, no pitch range expansion was found for the high-level tone in children or adults. These patterns suggested that 6-year-olds are adult-like in producing preboundary pitch cues that disambiguate compounds and lists. These findings are discussed in terms of the modulation of pitch information at the word and phrase levels and the role of pitch as a cue in response to boundaries in Mandarin.

Mandarin Tonal Perception in Question vs. Statement Sentences by Children with Cochlear Implants

Yanan Shen, Ivan Yuen and Ping Tang

Mandarin-speaking children with cochlear implants (CIs) face challenges in tonal perception since CI devices cannot transmit fundamental frequency (f0) effectively, though early implantation typically leads to better tonal perception. However, previous studies primarily examined children's perception of isolated tones. As sentence intonation, such as question vs. statement, can affect how tones are realized, it is unclear to what extent children with CIs also perceive tones in different sentential contexts and whether early implantation facilitates their tonal perception. Sixty 3-7-year-old children with CIs and 60 age-matched normal-hearing (NH) children were recruited. Their perception of tones in the final position of question and statement sentences was tested using a picture-pointing task. The results showed that while the NH group was equally accurate in tonal perception in both types of sentences, the CI group's perception accuracy of rising and dipping tones was significantly different between the two sentence types. Furthermore, early implantation was correlated with more accurate tonal perception in statements but not in questions. Therefore, children with CIs' tonal perception ability was contingent on intonation, and the effect of early implantation on tonal perception in sentences was restricted to statement sentences.

Cross-linguistic Influence on Intonation Acquisition: A Study on the Production of L2 Mandarin and L3 English Intonations by Uyghur Speakers

Tong Li and Hui Feng

Intonations in different languages serve the universal function of conveying communicative information and expressing affective meaning, while the prosodic encoding of the same intonational meaning is language-specific. For native speakers of Uyghur in China learning Mandarin as L2 and English as a third language (L3), few studies have been on their cross-linguistic acquisition of prosodic encoding of interrogative and declarative intonations in non-native languages.

Applying the L2 Intonation Learning theory (LILt), this study investigated the prosodic patterns of declarative and interrogative intonations in Uyghur, Mandarin, and English by 20 Uyghur speakers and compared them with 20 L1-Mandarin speakers and 10 L1-English speakers. Findings showed that the prosodic encoding of Mandarin intonation by Uyghur speakers was assimilated to Uyghur. When the pitch competition between lexical tones and interrogative intonation occurred, Uyghur speakers prioritized the prosodic encoding of intonation rather than the lexical tones, resulting in the gradual rising pitch level and rising boundary tones in Mandarin questions with the dipping tones (T3) and falling tones (T4). Regarding their English intonation, the delayed divergence between the downward declarative and upward interrogative intonations suggests that Uyghur speakers relied more on the sentence-final prosody in L1-like norms to distinguish the two intonations.

Peak alignment in Afro-Mexican Spanish: an exploratory analysis

Gilly Marchini

This paper presents a descriptive analysis of peak alignment in Afro-Mexican Spanish, a largely unexplored variety spoken in the South-west of Mexico. Sociolinguistic interview data was collected from one 51-year-old, female speaker, with a total of 122 broad focus, declarative Intonational Phrases annotated according to Sp_ToBI protocol.

Analysis reveals that whilst peaks align on the tonic syllable across open and closed syllables, there is an interaction with the nasality of the following sound: if present on the segmental string, peaks align on following nasals regardless of intervening syllable boundaries. In the case of closed syllables, i.e., with coda /n/, e.g., descendiente [de.sen̪.ˈdjen̪.te], peaks align tonically (90.5% of instances). For open syllables, i.e., with /n/ as following onset, e.g., mexicano [me.xi.ˈka.no], peaks align post-tonically (100% of instances).

Whilst tonic peak alignment is noted across Afro-Hispanic varieties, the role of the nasal is unexplored. Nor is it common in non-Afro Mexican Spanishes, where instead delayed peak alignment occurs. We consider the bearing of this upon the dialect-specific nature of the Segmental Anchoring Hypothesis (SAH), with reference to future experimentation required to test whether a lax SAH or articulatory phonological model can best account for these features.

Prosodic aspects of Brazilian L2 English: A comparison of duration-based rhythm and F0 measures with American English, Indian English, and Brazilian Portuguese

Leônidas Silva Jr., Jackciele Silva and Philipp Meer

This study investigates prosodic aspects of rhythm and intonation in the production of Brazilian L2 English (BrazE) in comparison to American English (AmE), Indian English (IndE), and Brazilian Portuguese (BP) based on prosodic-acoustic features of speech. Previous research suggests that IndE is more syllable-timed in terms of the syllable/stress-timing continuum than L1 varieties of English, such as AmE. The present paper hypothesizes that BrazE – due to L1 cross-linguistic influence of BP, which has been suggested to be syllable-timed – has lower variability of duration-based rhythm and F0 measures compared to AmE – just like IndE. We analyze several duration-based rhythm and F0 measures – suggested by recent L2 speech studies – in only-male read speech of BrazE, AmE, IndE, and BP (with five speakers per variety) using linear mixed-effects modeling. Results show significant differences between both BrazE-AmE and IndE-AmE in duration-based rhythm measures (%C, nPVI-V, Δ-S, YARD-S, RR-S, z-scored syllable duration, speech, and articulation rates) and F0 measures (z-scored F0, F0 semi-amplitude interquartile, F0 skewness). Minor, primarily non-significant differences were observed between BrazE and IndE/BP. Overall, with a view to duration-based rhythm and F0 measures, the results provide preliminary evidence that confirms our hypothesis that BrazE leans toward the syllable-timed pole of the rhythm continuum, not unlike IndE and BP.

Prosodic focus marking in Wa

Zenghui Liu and Min Wang

This study examines the prosodic realization of focus in Wa, a non-tone language in the Mon-Khmer language branch of the Austro-Asiatic language family. Using a semi-spontaneous approach, we elicited verb-subject-object (VSO) sentences from native Wa speakers (Awa dialect) in various focus conditions. We examined the use of prosodic cues, including duration, pitch span, pitch maximum, and pitch minimum, for encoding focus at the sentence-medial position. We found that Wa speakers varied prosodic cues for encoding focus types which differed in scope. Specifically, in narrow focus, Wa speakers not only shortened the duration of the focal constituent but also compressed the pitch span and raised the pitch maximum of the focal constituent compared to the same constituent in the broad focus condition. However, Wa speakers' use of prosodic cues for encoding narrow focus and focus types differing in contrastiveness was less consistent. These results demonstrate that pitch-related cues and duration vary for encoding focus types in Wa, presenting differences from existing studies.

Interpretation of single vs.~multiple wh-questions in semi-spontaneous Urdu

Farhat Jabeen

In languages with flexible word order, multiple wh-questions allow us to investigate the relationship between word order and prosodic phrasing. This study analyzes the prosodic phrasing of wh-questions in Urdu, a language with variable word order. We investigate if multiple wh-questions in Urdu pattern with single wh-questions or narrow focus in terms of word order and prosodic phrasing. We analyze word order variation in wh-questions, the prosody of in-situ wh-phrases compared with their scrambled counterparts, and the prosodic phrasing of wh-questions produced in semi-spontaneous speech. The GAMM analysis of time-normalized F0 showed that the intonation of wh-phrases differed in single and multiple questions. While no difference was found in the F0 of sentence initial wh-phrases in single and multiple questions, the F0 of wh-phrases scrambled to the sentence medial position in multiple questions was similar to that of narrow focus. Based on this, we provide a differential analysis of in-situ and scrambled wh-phrases in multiple wh-questions and compare it with the prosodic phrasing in single questions. Our data provides evidence for the interplay between word order and prosodic phrasing and contributes to the discussion regarding the analysis of wh-phrases as focused entities.

Speech Genre Classification in Online Multimedia Platforms: A Cross-Modal Approach Integrating Text and Prosody

Sin-Jhang Che and Alvin Cheng-Hsien Chen

The rise of online multimedia, particularly on YouTube, has transformed information dissemination. Specifically, this study examines the multimodal nature of these online speeches, focusing on two prevalent speech genres in Taiwan Mandarin YouTube content: entertaining and informative clips. We collected 100-minute video clips from sixteen influential YouTubers for each genre, and segmented the clips into inter-pause units (IPUs) for subsequent analyses. For each IPU, acoustic features describing durational, rhythmic, and pitch patterns were derived from its speech signals, while bag-of-word lexical features were developed from its textual content. Our objectives were twofold: firstly, to explore the genre-specific prosodic patterns using the proposed prosodic feature set, and secondly, to evaluate the additional contribution of these prosodic features in enhancing the accuracy of speech genre classification when integrated with textual features. Results show that the ensemble model outperforms prosody-only and text-only mono-modal models with an 84.6% accuracy, suggesting the complementary role of prosodic features in speech genre classification. Furthermore, our findings underscore the impact of semantic topics on textual features, potentially leading to misclassifications of topic-neutral IPUs in monomodal models. This study highlights the imperative consideration of both prosodic and textual features in determining speech genres within multimodal discourse.

CASS-AGING Corpus: The Development of Speech Prosody across the Mandarin-Speaking Adult Lifespan

Qian Li, Ziyu Xiong and Aijun Li

While studies have indicated a declination in the perception and processing of speech prosody due to normal aging, it is remained under-studied how elderly speakers might differ from the young adults in the production of prosody – especially for languages with interesting tone-intonation interactions – and how potential age-related changes might have been developed across the adult lifespan. This paper presents CASS-AGING, a speech corpus designed to examine the development of speech prosody across the Mandarin-speaking adult lifespan.

CASS-AGING was designed to include both read speech and spontaneous speech from 210 monolingual Beijing Mandarin speakers from 18-75 years old. For the read speech subset, each speaker was elicited with four lexical tones from 1156 monosyllabic words, 16 tonal combinations from 449 disyllabic words, 120-123 phonetically balanced utterances with varying lengths and sentence types (e.g., statements, questions), as well as a short discourse. For the spontaneous speech subset, each speaker was asked to complete a picture-description task and to give a monologue with hints.

In this paper, we aim to present our experiences in the corpus design and data collection, as well as the planned annotation procedures and possible analyses. Potential applications of the CASS-AGING corpus are also discussed.

Lexical Tone in Bilingual Crosstalk

Xin Wang and Bob McMurray

Spoken word recognition (SWR) is characterized by competition, as the lexical processor needs not only to interpret the unfolding speech input, but also to inhibit the activation of non-target candidates. This competition has been extended to investigations in bilingualism to understand how bilingual listeners recognize spoken words in one language that sound similar to words in the other. Lexical tones, has been shown to provide independent cues for lexical access within a tonal language. If tones are crucial in SWR, a key question is whether this linguistic knowledge is utilized in bilingual SWR. To address this question, we used the Visual World Paradigm due to its temporally sensitive measures of lexical activation and competition. Through two experiments, we investigated whether lexical tones provided independent cues in cross-language lexical competition, compared to segments. In Exp1, we presented natural native English tokens to Mandarin-English listeners; while in Exp2, we presented synthesized English tokens with superimposed Mandarin tones. We observed competition effects were larger when both segment and tone cues were available in the stimuli. These results first demonstrate the obligatory role of lexical tones in cross-language lexical competition in VWP.

Regional variation in pre-boundary lengthening from a horizontal and vertical perspective: Evidence from German dialect- and standard-targeted speech

Nadja Spina and Alfred Lameli

"Pre-boundary lengthening (PBL), the increase in duration of segments preceding a prosodic boundary, has been suggested to be a universal phenomenon that is implemented in language- specific ways. So far, research has focused on differences in the implementation of PBL across languages but never within languages. Simultaneously, PBL has mainly been investigated experimentally, while the need for analyzing speech corpora, not designed for studying PBL, rather than speech obtained in laboratory settings has been distinctly expressed.

The present study investigates regional variation in PBL in German, viewing the implementation of PBL from a horizontal perspective, i.e. between different German dialects, and from a vertical perspective, i.e. between each dialect and Standard German. Data from a corpus designed to investigate segmental characteristics of German dialects, containing dialect- and standard-targeted speech, were analyzed for PBL. Results reveal differences in PBL from a horizontal and vertical perspective. The dialect data show regional variation in PBL while the standard-targeted data approach PBL-patterns such as observed for Standard German."

Not only pitch: individual differences and priming of the implicit prosody of ambiguous only-association

Joy Mills and Sasha Calhoun

There has been increased interest in the effect of individual differences such as working memory and Autistic-like traits (AQ) on the processing of prosody, and specifically implicit prosody (prosody “heard” during silent reading). People with higher AQ scores are predicted to have lower sensitivity to pitch-related prominence. Implicit prosody can be measured using a bimodal priming paradigm where participants hear three auditory primes with the same prosodic disambiguation before a sentence with ambiguous association is presented visually, followed by a binary forced choice. Previous research has found a difference in processing based on AQ, but using relative clauses which are influenced by factors including rhythm, breaks and pitch-related focus. The current study uses pitch-based disambiguation of only-association. Our preliminary analysis shows effective priming: participants were more likely to choose the unexpected VP association after primes with a contrastive pitch accent on the verb. We also find, as predicted, that those with more autistic-like traits choose the unexpected verb-association less overall. This trend increases with higher working memory scores, suggesting that individuals who are more sensitive to pitch features must also be able to remember them, and that sensitivity alone is not enough to influence implicit prosody.

The impact of prosodic boundary and information structure on tonal coarticulation in spontaneous Cantonese

Xin Gao, Cesko Voeten and Mark Liberman

The current study examines the presence of PLR in spontaneous Hong Kong Cantonese using the CantoMap Map-Task dataset and explores the influences of prosodic boundaries and information structure (the givenness of a word in the discourse) using generalized additive mixed models (GAMMs). The findings confirm the presence of PLR in spontaneous speech. Moreover, the study demonstrates that PLR is stronger in the absence of a prosodic boundary, indicating the substantial influence of prosodic boundaries on tonal coarticulation. Furthermore, the study examines how the f0 of the following syllable impacts the f0 realization of the preceding syllable, revealing a correlation between the phonetic context and the realization of tonal coarticulation.

Can L2 speech rate surpass L1? Evidence from Mandarin learners of Japanese with and without immersion

Zhiqiang Zhu and Peggy Mok

This study investigates the intriguing scenario where L2 learners can outpace their L1 speech rate. Prior research indicates a faster speech rate in Japanese compared to Mandarin. However, the question remains whether native Mandarin learners can overcome their inherently slower L1 speech rate when speaking L2 Japanese. We assessed 15 N1-certified Mandarin learners of Japanese, divided by immersion experience—seven with at least a year in Japan, and eight without immersion. Their speech rates in both languages were measured against those of ten native speakers per language, including reading and spontaneous speech.

Challenging the L1 superiority belief, our findings reveal that the immersed group could match the faster speech rates of native Japanese speakers, thereby exceeding the speech rates of their L1 Mandarin, which are similar to other native Mandarin speakers. Conversely, the non-immersed group’s Japanese speech rate was comparable to or slower than their L1 Mandarin.

Subsequent analysis probed how speech rate correlated with other learner variables such as gender, age, study duration of Japanese, and length of residency in Japan. The findings also highlight immersion as the critical factor of speech rate. This study extends our knowledge of bilingual fluency, providing new perspectives on L2 prosody mastery.

Tonal patterns of the Mandarin Third Tone Sandhi produced by Japanese-speaking L2 learners

Tong Shu, Zhiqiang Zhu and Peggy Mok

While extensive research has been conducted on the L2 perception and production of Mandarin lexical tones, the higher prosodic patterns, such as tone sandhi, remain less explored. This study examined the L2 production of the Mandarin Third Tone Sandhi (T3 Sandhi) by Japanese speakers at two Mandarin proficiency levels (intermediate and advanced). The participants read disyllabic stimuli with all possible tonal combinations of the T3 Sandhi. Different from the common approach which mainly relied on native speakers' categorization of L2 learners' tone production, we adopted a data-driven approach using hierarchical clustering to identify the distinct tonal patterns for each T3 Sandhi combination within each group.

The results revealed a complex interplay of various factors influencing L2 production of the Mandarin T3 Sandhi, such as L1 Japanese pitch accent patterns, phonetic motivation of different T3 Sandhi, and L2 Mandarin tone inventory. The suspected influence from L1 Japanese pitch accent patterns is noted in intermediate-level learners, but advanced learners can overcome such influence. In both L2 learner groups, we found over-generalization of T3 Sandhi. In general, our study showed the transfer of L1 phonological processing to L2 tone sandhi production at an earlier stage of L2 acquisition.

The Role of Auditory and Visual Modality in Perception of English Statements and Echoic Question by Chinese EFL Learners

Shanpeng Li, Yinuo Wang, Shifeng Xia, Zhiqiang Tang, Ping Tang and Yan Feng

Previous research underscored the role of auditory and visual cues in perceiving statements and questions, yet with conflicting conclusions regarding their relative significance. It was argued that English speakers relying predominantly on auditory cues, with limited impact from visual cues for intonation. Given the variability observed in language-specific utilization of auditory and visual modalities for interpreting statements and questions, as evidenced in studies involving Dutch and Catalan, the generalizability of findings to other languages remains uncertain. Therefore, the study aimed to investigate the influence of auditory and visual cues on the perception of English statements and questions among Chinese EFL learners.

A total of 56 Chinese EFL learners participated in the audiovisual perception study, categorized into three blocks: audio-only (AO), visual-only (VO), and audiovisual (AV) conditions. Additionally, to explore the contribution of specific facial areas, the VO and AV conditions were subdivided into full-face, upper-face only, and lower-face only conditions.

The result revealed that Chinese EFL learners lean towards visual cues, particularly upper face when perceiving English intonations. This inclination could be attributed to factors such as whispered condition and cultural inclinations. Recognizing and incorporating these influences into teaching approaches can significantly enhance the comprehension of intonation among EFL learners.

Robust evaluation metrics for automatic speech rate computation

Mireia Farrús, Wendy Elvira-García and Juan María Garrido-Almiñana

Speech rate is an essential element for prosodic analysis. However, its acoustic measurement for large-scale applications can be hard if attempted by manual means. Therefore, automatic computation of speech rate is a good alternative provided that a reliable enough performance is guaranteed. In this light, the assessment of the performance of speech rate estimation tools has been attempted using several metrics, being correlation coefficient one of the most widely used. Nevertheless, it is not clear to what extent these methods offer a reliable enough measurement of the performance of these automatic systems. To address this issue, the current paper reviews the different evaluation methods that have been used according to the literature to assess the automatic computation of speech rate, and tests them on a corpus of read and spontaneous speech in Spanish. The obtained results show that error-based metrics are more robust and appropriate than correlation coefficients. Based on the empirical results, the current study concludes with a proposal of standard measures for evaluating automatic speech rate computation.

The critical rhythm measures in classifying and assessing L2 Chinese speech

Sichang Gao and Chao Kong

This study aims to identify the rhythmic measures that present challenges for L2 Chinese learners and whether these challenges are more unique for learners with specific L1 backgrounds. Also, we analyzed the role of these rhythm features in assessing the perceived naturalness of L2 Chinese speech. 107 L2 Chinese learners from five L1 backgrounds, including Vietnamese, Japanese, Thai, Russian, and Korean, were recruited to elicit L2 speech using an identical reading task. Perceived naturalness was evaluated based on the subjective judgment of six Chinese native speakers. Linear discriminant analysis, combined with linear model analysis, showed that consonant-related measures (VarcoC, nPVI-C, ΔC, rPVI-C) were better indices for classifying L2 Chinese rhythm by five L1-background groups, indicating L1 influence on L2 rhythm. However, the effect of consonant-related measures was mainly demonstrated in Japanese and Russian learner groups. Vowel-related rhythm measures (ΔV, VarcoV, rPVI-V, nPVI-V) showed a tendency toward universal constraints, as all L2 groups manifested mastery in the rate-normalized variability of vowel duration (VarcoV, nPVI-V) and the deficiency in the raw indices (ΔV, rPVI-V). Linear regression analysis also revealed that vowel duration (meanV) and the complexity of consonants (ΔC) contributed to the perceived impression of naturalness in L2 Chinese.

An automatic prosodic transcriber for the P-ToBI system

Wendy Elvira-García, Marisa Cruz, Marina Vigário and Sónia Frota

This study introduces a rule-based Praat script designed to generate P-ToBI labels based on the pitch contour given a tier with by-syllable intervals and stress marks.

The system was trained on a 96-sentence corpus comprising all Nuclear Pitch Accents (NPA) and Boundary Tones (BT) in European Portuguese (EP). Evaluation was conducted on a separate corpus of 146 sentences showing a success rate of 73.8% (k=0.6) for NPA and 78.7% for BT (k=0.6). The qualitative analysis of errors, excluding those stemming from the pitch tracking algorithm, exposes challenges in accurately identifying falling NPAs, particularly instances of L*, H*+L, and H+L* followed by a low BT (although they can be accurately distinguished using an additive model). The performance of the system contrasts with results obtained with similar procedures for other Romance languages that get to 90% of success.

We argue that the performance difference stems from principles underlying different ToBI systems (with P-ToBI being more phonological), and specificities of the phonological system of EP, namely word-final vowel reduction and deletion. This suggests that a rule-based approach relying solely on the acoustic signal may not be the most suitable for European Portuguese.

L2 comprehension of focus-to-prosody mapping by Mandarin learners of English

Aishu Chen and Haoyan Ge

This study investigates how Mandarin learners of English use prosody to comprehend focus in English sentences with the focus particle “only” and how L2 proficiency affects their performance.

We adopted the paradigm of Ge et al. (2020) and conducted a comprehension experiment on two groups of L2 learners at two levels of English proficiency (advanced: N=36, intermediate: N=30). Participants were presented with question-answer dialogues and were asked to judge whether the answer (with appropriate OR inappropriate prosody) made sense for the question based on a given story as quickly as possible.

The results showed that appropriate prosody triggered more “YES” judgements than inappropriate prosody in both groups of L2 learners, regardless of their L2 proficiency. Besides, L2 learners were faster in the appropriate-prosody condition than in the inappropriate-prosody condition in comprehending focus. It seems that Mandarin learners of English match native speakers of English in making use of prosody to interpret focus. Significant differences were observed between two groups of L2 learners: advanced L2 learners gave more “YES” responses and were much faster than intermediate L2 learners with appropriate focus-to-prosody mapping. Our results provide further evidence that L2 proficiency plays a role in L2 comprehension.

A Study of the Sensitivity of Subjective Listening Tests to Inter-sentence Pause Durations in English Speech

Paul Owoicho, Joshua Camp and Tom Kenter

Inter-sentence pauses are silences occurring between sentences in a paragraph or dialogue. They are an important aspect of long-form speech prosody, as they can affect the naturalness and effectiveness of communication. When evaluating the output of long-form speech synthesis systems, it is crucial to understand the sensitivity of commonly used tests to variations in inter-sentence pause durations, as this sensitivity impacts the usefulness of such evaluations. However, perception of inter-sentence pauses in long-form speech synthesis is not well understood. Previous work often evaluates pause modelling in conjunction with other prosodic features making it hard to explicitly study how differences in inter-sentence pause lengths are perceived. To fill this gap, we investigate the sensitivity of subjective listening tests to changes to the durations of inter-sentence pauses in long-form speech, by comparing ground truth audio samples with renditions that have manipulated pause durations. Using multiple datasets to cover a variety of domains, we find that listening tests are not sensitive to variations in pause lengths unless these deviate from the norm exceedingly. Our evaluation experiments in this study can be considered preliminary work, the findings of which will have implications for evaluation experiments run on actual synthesized long-form speech.

A Cross-linguistic Study on Audiovisual Perception of Prosodic Prominence by Chinese and English Observers

Ran Bi and Marc Swerts

Speakers and their conversation partners use both auditory and visual cues (facial expression) to highlight important information by making some words more prominent. Prior work on languages like Dutch and English has shown that intonation markers such as pitch accents, duration and loudness as well as facial expression such as head, eyebrow and mouth movements are often exploited to signal and interpret prosodic prominence. However, little is known about how observers of different linguistic backgrounds (Chinese and English) perceive prominence of these languages, and how audiovisual cues affect prominence perception of the two languages, especially when audiovisual cues are incongruent. Using naturally elicited stimuli from Chinese and English speakers, a perceptual experiment was conducted to measure both L1 and L2 observers' reaction time and accuracy in a task of judging which word was prominent in speech. The observers were exposed to both Chinese and English stimuli in three different formats: audio-only, audiovisually congruent, and audiovisually incongruent. Results revealed that (1) visual cues were important for prominence perception especially in incongruent stimuli; (2) observers of both languages identify prominence more easily and accurately in Chinese than in English; (3) there are consistent correlation between reaction time and accuracy.

Carryover Tonal Variations for Speech Recognition in Standard Chinese

Hana Nurul Hasanah, Qing Yang and Yiya Chen

Substantial pitch variation due to tonal coarticulation occurs when lexical tones are produced in succession. When a coarticulated tone is excised out of context and played in isolation, coarticulatory pitch variations may inhibit tone recognition. It remains unclear how listeners utilize coarticulatory pitch cues for online speech recognition in context. Using the printed-word eye-tracking paradigm, we tested the recognition of the high tone in a low-high tonal sequence by native Standard Chinese (SC) listeners. In this sequence, the high tone exhibits coarticulatory rising f0. We manipulated the presentation of the preceding low tone: auditorily and visually present, auditorily absent (i.e., substituted by pink noise) but visually present or visually replaced by a high tone (to prompt inappropriate tonal coarticulatory cue). Analyses of the point of divergence and proportions of eye fixations revealed that listeners' correct fixations at the high-tone target started early and increased quickly even though only the rising f0 part of the high tone was played auditorily, with a gradual delay following the compatibility between the visual and auditory stimulus presentations. The immediate utilization of tonal coarticulation for speech recognition by SC listeners suggests the need for fine-grained coarticulatory information in speech representation and processing.

Is Pitch Contour Sufficient to Encode Prosody in Neural Text-to-Speech?

Alex Peiró Lilja and Mireia Farrús Cabeceran

Nowadays speech synthesis has reached levels of voice quality and naturalness close to human. This has been achieved thanks to the rapid evolution of generative architectures deployed for neural text-to-speech (TTS). Many approaches have been proposed to encode speech style --i.e. prosody attributes-- leveraging these models in order to transfer it to the generated speech. The most common acoustic features for this purpose are the spectrograms. However, is the whole frequency representation really necessary to learn speech attributes? To answer this question, in this work we propose the sparse pitch matrix (SPM), an sparse and binary representation of the pitch sub band. We assumed that pitch is sufficient to make the model extrapolate the rest of the prosody aspects. To study its impact, we performed an experiment built upon the unsupervised global style tokens conditioning the Tacotron2 decoding. The tokens were fed with the encoded SPMs during training, similarly to the original approach. From the posterior analysis we found that: 1) there are significant differences in many prosody attributes between tokens, and 2) all tokens, in isolation, provide acceptable levels of quality, intelligibility and naturalness, according to human evaluators.

Talker-specific perceptual learning about lexical stress: stability over time

Giulio G.A. Severijnen, Verena M. Gärtner, Runa F.E. Walther and James M. McQueen

Talkers vary in how they speak, resulting in acoustic variability in segments and prosody. Previous studies showed that listeners deal with segmental variability through perceptual learning and that these learning effects are stable over time. The present study examined whether this is also true for lexical stress variability. Listeners heard Dutch minimal pairs (e.g., VOORnaam vs. voorNAAM, ‘first name’ vs. ‘respectable’) spoken by two talkers. Half of the participants heard Talker 1 using only F0 to signal lexical stress and Talker 2 using only intensity. The other half heard the reverse. After a learning phase, participants were tested on words spoken by these talkers with conflicting stress cues (‘mixed items’; e.g., Talker 1 saying voornaam with F0 signaling initial stress and intensity signaling final stress). We found that, despite the conflicting cues, listeners perceived these items following what they had learned. For example, participants hearing the example mixed item described above who had learned that Talker 1 used F0 perceived initial stress (VOORnaam) but those who had learned that Talker 1 used intensity perceived final stress (voorNAAM). Crucially, this result was still present in a delayed test phase, showing that talker-specific learning about lexical stress is stable over time.

Prosodic correlates of negative rhetorical questions in Lombard Italian

Laura Colantoni, Michela Ippolito and Mariapaola D'Imperio

Our goal was to investigate whether Lombard Italian speakers reliably use prosodic features to distinguish canonical wh-questions (CQs) and non-canonical rhetorical wh-questions (RhQs). We hypothesized that such speakers could identify each question type by using prosody alone, and we designed both a perception and a production task to test this hypothesis. The former consisted of a forced-choice identification task with 32 target stimuli. To elicit the production data, participants performed a discourse completion task with the same number of stimuli as in the perception study. Utterances were intonationally labeled and acoustically analyzed for pitch change (prenuclear pitch accent and nuclear contour), initial and final pitch, and relative duration. Results revealed that participants were highly accurate at identifying RhQs but less so at identifying CQs. Results of successful productions showed that initial but also final cues were systematically used, with pitch level being lower in RhQs than in CQs, and final boundary tones reliably differentiating RhQs from CQs. Finally, RhQs were significantly longer than CQs. We conclude that speakers do rely on both intonation and duration cues to distinguish CQs from RhQs.

The prosody of Italian newsreading: a diachronic analysis

Michelina Savino, Simon Wehrle and Martine Grice

Current research in the field has identified a broadcast newsreading style and described it in terms of variations in prosodic parameters in comparison to, for example, a narrative style or conversational speech. However, less is known about how the prosodic style of newsreading has evolved diachronically. With the aim of filling this gap, we compared the prosodic style of Italian newsreading originally recorded at the end of the 60s with the productions of two newscasters in 2005, who were instructed to read the '60s news using their current newscast reading style. To analyse the intonation of these two styles, we adopted an innovative methodology capturing the time-varying dynamics of F0 along the dimension of Wiggliness (time-varying F0 in terms of slope changes), and Spaciousness (F0 excursions at largest rises and falls). Results show that, although in terms of Wiggliness and Spaciousness the intonation styles were similar across the two eras, mean pitch was higher in the modern era. Additionally, there was a considerable increase in speech rate. This increase in rate was mainly achieved by reducing the number and duration of pauses, thus producing much longer inter-pausal units as compared to the newsreading style in the 60s.

Immediate sentence repetition in autism: Effects of listening background, mode of presentation, and semantic content

Chen Zhao, Qingqi Hou, Ariadne Loutrari, Li Wang, Cunmei Jiang and Fang Liu

Autism has been linked to various speech and language deficits, including difficulties in immediate sentence repetition. In this study, we adopted the sentence repetition paradigm to investigate the effects of mode of presentation (speech versus song), semantic content (news-like, story-like, and non-semantic), and listening background (quiet versus noisy), on immediate sentence repetition in a group of 30 Mandarin-speaking autistic children and 29 non-autistic children. Results showed that repetition accuracy was poorer in the autistic group, although the group difference may be partially the product of variation in verbal ability and age. Both groups performed worse for sung as opposed to spoken sentences, and for news-like and non-semantic relative to story-like sentences. We discuss how factoring in multiple variables, including individual differences, can further our understanding of sentence repetition ability in autism.

Phrase-Final Voice Quality Variation Among Black and Latinx Southern California Youth

Nicole Holliday

In the U.S., differences between ethnolectal varieties such as Latinx English and African American English have been well-described for decades, especially in the realm of segmental phonological variation. However, few studies have examined patterns of voice quality features across ethnolectal varieties of U.S. English. The current study examines acoustic correlates of creakiness, modalness and breathiness (H1-H2, HNR, and CPP values) in the speech of 24 Southern California high school students: 17 Latinx and 7 Black American. Results of multiple regression analyses reveal significant differences in phrase-final voice quality, with Latinx students displaying patterns that are breathier and Black students displaying ones that are more modal or creakier. These results may be due to different strategies across ethnolects for accomplishing acoustic cues to phrase-final position. They can also reflect different levels of participation in the ongoing increase in the use of creaky voice among young speakers of Mainstream American English. This study acts as a first step towards understanding how racialized communities in the U.S. may participate in sociophonetic changes in progress that interact with the making of prosodic meaning.

Establishing the domain of intonation patterns: syllabic nuclei vs syllabic rimes

Michaela Svatošová and Jan Volín

Descriptions of intonation patterns are often based on the values of the fundamental frequency in vowels. In Czech, however, syllabic rimes can contain sonorous consonants and their relevance in intonation is unclear. The present study examined whether F0 contours in phrase-final syllables with different types of coda (no coda, sonorous coda, obstruent coda) are more similar in the domain of syllabic nuclei or in the domain of syllabic rimes. The material consisted of a two-hour collection of audiobook samples recorded by sixteen professional actors, providing 3756 phrase-final syllables. The contours were analysed first simply in terms of their realisational space and then in more detail with functional principal component analysis and Legendre polynomials.

The results indicated that the domain relevant for the production of intonation patterns is the sonorous part of the syllabic rime, which showed comparable values of F0 range across all syllable types. Typical contours had straight shapes; however, some contours were considerably curved, which led to substantially different shapes and F0 ranges in the syllabic rimes compared to the nuclei for syllables with a sonorous coda. The present findings provide a basis for perception experiments in the domain of intonation patterns in Czech.

Sensorimotor influences on infant speech perception also target prosody

Sónia Frota, Cátia Severino, Jovana Pejovic and Marina Vigário

Sensorimotor influences on infant auditory speech perception have been shown to modulate phoneme discrimination, highlighting the integration of sensorimotor information and auditory speech prior to production. Although the perception of prosody is multimodal, sensorimotor influences on prosody perception have not been investigated. We examined the generalizability and potential cross-domain impact of sensorimotor influences beyond phoneme perception, by testing a novel phoneme contrast (/da/-/za/), a stress contrast (trochaic/iambic) and an intonation contrast (H+L* L% / H+L* LH%). In 3 experiments, 6-month-old infants listened to the speech contrasts without and with oral-motor impairments induced by two teething toys: a gummy teether and a flat teether. Infants discriminated the segmental contrast only in the absence of a teething toy. Similarly, the stress contrast was discriminated without a teether, while both teethers disrupted discrimination. Discrimination of the intonation contrast was found both in the absence of the teething toys and in the gummy teether condition, but not in the flat teether condition. Our results are the first to suggest that sensorimotor influences modulate speech perception beyond phoneme discrimination. Their effects on infants' prosody perception were different for stress and intonation, pointing to specificities in the perception-production link with stress being more disrupted than intonation.

Breathing features and their impact on speech perception of COVID-19 patients

Xiaoming Jiang, Lixin Yu, Leinuo Dai, Jinyang Chen and Zheng Yuan

Breathing features are valuable tools for detecting and diagnosing respiratory diseases from speech. The current study analyzed speech samples of interviews from 23 COVID-19 patients via social media platforms. Breathing features were extracted for each breath group (BG) locally and for each discourse globally. Perceptual tasks based on written texts and audio samples were conducted at both breath-group and discourse levels. PCA reduced breathing features to 1) at breath-group level: BG length, proportion of pauses, and speech rate; 2) at discourse level: discourse length, average BG duration, average inter-breath-group pauses (IBP) duration, and proportion of IBP to BG each within a discourse. Linear models showed that at breath-group level, speaker's BG length can be positively predicted by the perceived text valence and negatively by the perceived text fluency, while proportion of pauses within BG negatively by fluency. At discourse level, the average BG duration has a negative predictive effect on the perceived probability of illness; additionally, the longer BG duration predicts higher illness severity, while the higher proportion of IBP to BG predicts lower illness severity. Our study presents a data-driven approach of breathing features associated with respiratory diseases and demonstrates the way how these features interact with speech perception.

Some prosodic consequences of varied discourse functions in a Cantonese sentence-final particle

Jonathan Him Nok Lee, Ka-Fai Yip, Mark Liberman and Jianjing Kuang

This study investigates the prosodic correlates of varied discourse functions for a Cantonese sentence-final particle. Three functions of the particle ge2 were examined: blaming others (Blame), defending oneself (Defend), and asking for reasons (Reason). Ten native adult speakers of Cantonese participated in the production experiments. Results of both Smoothing Spline ANOVA and Generalized Additive Mixed Models suggest that, despite the same citation tone of ge2 (high-rising tone), the pitch of Reason is significantly lower than that of Blame and Defend, and there is no significant difference between the latter two. Besides, mixed-effects regression shows that the relative vowel duration of Defend is significantly shorter than the other two functions. Additionally, k-means clustering suggests that Reason can be reliably classified based on its lower pitch. Blame and Defend have similar pitches and primarily differ in relative vowel duration. Our results suggest that different functions of the same Cantonese particle have different phonetic realization in pitch and duration.

Production of non-native quantity contrasts by native speakers of Cantonese, English, French, and Japanese

Albert Lee, Yasuaki Shinohara, Faith Chiu and Tsz Ching Mut

In this study we compared the duration ratios of native speakers of Cantonese, English, French, and Japanese who produced non-native phonemic quantity contrasts in Japanese (two-way) and Estonian (three-way). These four L1 backgrounds differ in terms of the extent to which duration is used to mark quantity contrasts (e.g. short vs. long), ranging from non-phonemic (i.e. French) to systematic two-way (i.e. Japanese). A shadowing task was used to elicit participants' production. Estonian and Japanese stimuli (N = 360) were played in two separate blocks. The participants wore a pair of earphones in a quiet room and repeated the word they heard. The results showed that all participant groups were able to tell apart the quantity conditions, though for the Estonian target words Short vs. Long were better differentiated than Long vs. SuperL. The unexpectedly good performance of the French speakers in discrimination and identification (Lee et al., 2023) was replicated, as was the relatively poor performance of the Cantonese speakers. The theoretical implications of these findings are discussed.

Prosodic Decoding Profiles of Chinese Mandarin-Speaking Children under Visual and Audio Modality

Jue Yu, Kexin Zhang and Shiyi Zhu

The present study aimed to obtain a comprehensive understanding about how Chinese Mandarin-speaking children performed prosodic decoding of syntactic ambiguity under the visual and audio modality and to what extent their manipulation of implicit prosody differed from explicit one. Altogether 31 Mandarin-speaking children were recruited and required to resolve different types of syntactic ambiguity when reading and perceiving a sentence. Accuracy and reaction time in the visual and audio modes were measured and analyzed. The results showed: first, comparatively most children were better at applying explicit prosody to resolve most syntactic ambiguity than implicit one, though a positive correlation was found between the two. Secondly, syntactic ambiguity with an attachment structure increased the processing cost for most children, i.e. they spent more time on prosodic decoding, which agreed with Derivational Complexity Hypothesis. Finally, regardless of the syntactic ambiguity types, most children consistently held a bias to decode the pause at earlier-occurring prosodic boundaries under both audio and visual modality, which offered an interesting perspective from Late Closure Strategy.

Crossing Boundaries: Prosodic Aspects of Code-Switching Effects between Mandarin and English

Yao-Zhen Zeng

This study investigated how code-switching from Mandarin to English affected pitch range and pitch reset. Our intuition was that speakers tend to raise their pitch when producing English words embedded in Mandarin utterances. Utterances in Mandarin-only, English-only, and code-switched conditions were created to observe the acoustic distinctions. Two primary aspects were explored. First, we compared the pitch range values of keywords between the English-only and code-switched conditions. Second, we examined the manifestation of pitch reset patterns in keywords by comparing the Mandarin-only condition with the code-switched condition. The results show that in the keywords, the pitch range values in the code-switched condition significantly surpassed those in the English-only condition, suggesting a notable pitch difference in how the keywords were produced during code-switching. As for pitch reset, the values of keywords in the code-switched condition were significantly greater than those in the Mandarin-only condition, highlighting a discernible language transition (from Mandarin to English) in the code-switching process.

Tone acquisition in Chinese-speaking children: Developmental data of tone acceptability and contour pattern

Shu-Chuan Tseng

This paper presents developmental data from 798 Chinese-speaking children aged three to six. An overall assessment of tone acquisition is reported with descriptive results from tone acceptability judgement, pitch contour type labeling and F0 slope patterns. The order of tone acquisition is Tone 4, Tone 1, Sandhi Tone 3, Tone 3, and Tone 2. High-register tones are generally acquired before low-register tones. Pitch contour types of perceptually acceptable tones primarily conform to phonologically predicted tone contours including Sandhi Tone 3. Unacceptable tones are actually more diverse. Based on the F0 property, the proportions of immediately erroneous slope patterns in Tone 2 are higher than in Sandhi Tone 3. As a whole, our developmental data consistently suggests that Tone 2, instead of Sandhi Tone 3, is the most challenging tone to acquire for Chinese-speaking children.

Register as a motivation for change: a case of High Vowel Fricativization in Changzhou Chinese

Aixin Yuan and Jason Brown

Fricative vowels are unusual in languages worldwide. High Vowel Fricativization (HVF) is a model for the emergence of such sounds, which requires vowels to have consistent turbulence to generate wall noise source as the driving force. However, its phonetic foundation remains unclear.

To investigate, 38 native speakers of Changzhou Chinese produced a series of CV monosyllabic words with alveolar fricative vowel /ʑ̩/ and its plain counterpart, high front vowel /i/ as the nucleus across five tones. We measured the spectrograms, Harmonics-to-Noise Ratio (HNR) and several phonation measurements for each token.

Results show that the high vowel in lower register exhibits breathy phonation and HNR values similar to fricative vowels. It suggests that lower register /i/ is produced with audible turbulence, acting as one of the prerequisites for fricativization. Listeners may misperceive the vowel as a new sound category primarily cued by frication noise, eventually spreading the sound to other tonal categories and giving rise to a fricative vowel /ʑ̩/. Overall, this study provides the phonetic foundation for the emergence of frication noise in fricative vowels, arguing that register can be a motivation for change in the fricativization of high vowels.

Inconsistent prosodies more severely impair speaker discrimination of Artificial-Intelligence-cloned than human talkers

Wenjun Chen, Xiaoming Jiang, Jingyi Ge, Shuwan Shan, Siyuan Zou and Yiyang Ding

AI algorithms designed to clone human speaker identity are reportedly capable of replicating human-specific vocal confidence. However, whether listeners can accurately identify a single speaker expressing varying emotive states as the same individual remains unclear, particularly never in AI-to-AI pairings. This study asked thirty-six Chinese participants to judge whether identical speakers delivered pairs of Chinese sentences with incongruent or congruent prosody in human-only and AI-only scenarios. We found a marked decrease in the accuracy of identifying the same speaker under inconsistent prosody conditions compared to consistent ones, a trend evident in both human-to-human and AI-to-AI pairs. Meanwhile, correctly distinguishing between two speakers was more challenging than identifying a single speaker, with AI pairs reporting notably poorer performance than human-human pairs. We observed that listeners slowed down reaction times when faced with inconsistent prosody in the one-speaker scenario, whereas they reacted faster in two-speaker setups. Listeners reacted similarly fast in human and AI trials. Our findings suggest vocal prosodies can lead to within-speaker identity variation but around the average-based representations, which listeners can overcome and still recognise the same speaker across prosodies. Our results about speaker discrimination in AI voices also provide supportive evidence for the “out-group homogeneity effect”.

The effect of phonotactic constraints on tone sandhi application: A cross-sectional study of Xiamen Min

Chunyu Ge and Peggy P.K. Mok

Nonce-probe test has been extensively used to investigate the productivity of tone sandhi. The nonce words used in previous studies on Xiamen tone sandhi were usually disyllabic wug words with accidental-gap syllables. This study aims at isolating the effect of phonotactic constraints by investigating the application of Xiamen tone sandhi to both (1) disyllabic semi-wug words made up of real syllables and (2) disyllabic wug words consisting of one accidental-gap syllable and one real syllable. Picture-naming tasks were used to elicit the production of these conditions from children, teenagers, middle-aged adults and older speakers. The results showed that Xiamen tone sandhi was highly productive for semi-wug words but far less productive for wug words. Children and teenagers made some errors in applying the correct tone sandhi rules to real and semi-wug words, while their accuracy of applying tone sandhi to wug words was very similar to those of the middle-aged and older speakers. It is concluded that Xiamen tone sandhi is highly productive in phonotactically well-formed real syllables but less productive in phonotactically ill-formed syllables.

The intonation of Kriol: a first approach

Gabriela Braga, Sónia Frota and Flaviane Svartman

Kriol is the main language spoken in Guinea-Bissau, a multilingual country where the only official language is Portuguese. This study is a first description of the intonation of Kriol, focusing on the intonational contour of statements and lists, using the Autosegmental-Metrical framework. Although there has been an increasing interest in the prosody of varieties of Portuguese, there is still a lot to investigate about the prosody of languages genetically related to the colonizer language, like Creole languages. The analysis was developed based on semi-spontaneous speech data collected via a Discourse Completion Task and a Story Telling Task. The results showed that the intonation of statements in Kriol consists of a sequence of high pitch accents (H* or !H*) associated with the heads of phonological phrases, in a stepwise fashion, ending with a falling or low nuclear contour (H+L* L% or L* L%). Furthermore, because Kriol is a tense-moodaspect system language, phonological phrases were mapped to only one prosodic word. List intonation was characterized by varying contours across and within speakers, with rising contours predominating (L(H)H%). Overall, differently from European Portuguese (EP), Kriol shows high tonal density in statement intonation, and, similarly to EP, weak macro-rhythm, given the low alternation of high and low tones.

Interactive Prosodic Encoding of Tone, Focus and Sentence Type in Changli-Town Mandarin

Mengxue Cao, Tianxin Zheng, Hongna Li and Aijun Li

There is increasing evidence that many Chinese dialects, in comparison to Beijing Mandarin, exhibit distinct focus and intonation patterns. This study investigates how tone, focus and sentence-type-related intonation contribute to shaping F0 contours in Changli-Town Mandarin. Acoustic analysis of short sentences, produced by seven native speakers, was conducted, where sentence-final focus and declarative/interrogative intonation were considered as variables. The results reveal that (1) sentence-type-related intonation regulates the F0 curve through compulsory boundary targets, indicating a low target for declarative sentences and a high target for interrogative sentences, where incompatible tonal targets must be adjusted accordingly; (2) sentence-final focus modulates the global shape of the F0 curve by expanding the F0 range of on-focus syllables and oft-elevating the F0 register of all pre-focus syllables, where the pre-focus pitch elevation is argued primarily as a co-effect of focus and pragmatic purposes, enhancing listeners' attention to key information; (3) neutral tone emerges as the most common agent serving to coordinate tonal and intonational targets. Those findings suggest that in Changli-Town Mandarin, tone, focus and intonation interact to encode the F0 curve of a sentence by coordinating respective targets, with the boundary target exerting dominance. Those insights hold implications for prosodic typology.

The more complex the better? Mandarin tone perception by Cantonese and Hakka speakers

Siyi Lian and Min Liu

Much research has been conducted to investigate how native tonal language experience, in comparison with non-tonal language experience, shapes the perception of non-native tones. Little is known about how tonal experience of different dialects within a tonal language affects tone perception of the standard variety. The present study aimed to investigate how the complexity of tonal system in two Chinese varieties (i.e., Cantonese and Hakka), and how the tonal correspondences between each of the two varieties and Mandarin (i.e., the standard variety), affect Mandarin tone perception by speakers of the two varieties. The tonal system of Cantonese is more complex than that of Hakka, in terms of both tonal inventories and tonal categories. However, the correspondences between Cantonese tones and the Mandarin level/falling tones are looser than those between Hakka and Mandarin tones. An identification experiment and a discrimination experiment of Mandarin level and falling tones were conducted among native speakers of Guangzhou Cantonese, Meizhou Hakka (both with high level of Mandarin) and Mandarin. Results showed that the more complex the native tonal system, the better perception of Mandarin tones. Surprisingly, the tonal correspondences between the language variety and Mandarin did not seem to affect Mandarin tone perception as expected.

Human Vocal Attractiveness in British English as Perceived by Chinese University Students

Yifan Yang and Yi Xu

"Vocal attractiveness, as an important indicator of personal traits, was hardly explored from the perspective of foreign listeners. In this study, both English native speakers and Chinese university students were asked to evaluate synthetic English utterances with manipulated voice quality, formant dispersion, pitch shift and pitch range in same-sex and opposite-sex contexts.

While some deviant features are shown in the results, English and Chinese subjects followed the principle of body size projection to varying degrees, with preferences for breathiness, higher pitch of female voices to signal a small body size, and narrower formant distribution of male voices to signal a large body size. Breathiness was also preferred for male voices to reduce implied aggressiveness by other body-size indicators. However, noteworthy differences between the two groups existed. Overall, Chinese subjects gave higher mean ratings to both genders and demonstrated weakened dimorphic characteristics of preferences compared to English subjects. Furthermore, for the same-sex voices, English women and Chinese men provided significantly lower ratings than their opposite gender within the same group.

The cross-linguistic differences in perceived vocal attractiveness shown in these results could be due to various linguistic, cultural, psychological, and educational factors."

Interindividual variation in weighting prosodic and semantic cues during sentence comprehension – a partial replication of Van der Burght et al. (2021)

Constantijn L. van der Burght and Antje S. Meyer

Contrastive pitch accents can mark sentence elements occupying parallel roles. In “Mary kissed John, not Peter”, a pitch accent on Mary or John cues the implied syntactic role of Peter. Van der Burght, Friederici, Goucha, and Hartwigsen (2021) showed that listeners can build expectations concerning syntactic and semantic properties of upcoming words, derived from pitch accent information they heard previously. To further explore these expectations, we attempted a partial replication of the original German study in Dutch. In the experimental sentences “Yesterday, the police officer arrested the thief, not the inspector/murderer”, a pitch accent on subject or object cued the subject/object role of the ellipsis clause. Contrasting elements were additionally cued by the thematic role typicality of the nouns. Participants listened to sentences in which the ellipsis clause was omitted and selected the most plausible sentence-final noun (presented visually) via button press. Replicating the original study results, listeners based their sentence-final preference on the pitch accent information available in the sentence. However, as in the original study, individual differences between listeners were found, with some following prosodic information and others relying on a structural bias. The results complement the literature on ellipsis resolution and on interindividual variability in cue weighting.

The Effects of Autistic Traits on Pitch-Semantic Integration Processing: Evidence from an ERP Study

Li Xia, Ting Wang and Yuhan Jiang

Autistic traits (ATs) are various primary symptoms associated with autism spectrum disorders (ASD) and are continuously distributed within the general population. Previous studies from non-tonal language backgrounds have demonstrated that autistic individuals show enhanced pitch perceptual abilities, yet encounter challenges in higher-level semantic integration involving pitch. Considering the comorbidity in autistic individuals, to clarify the underlying causes, whether neurotypical people with high autistic traits share similar challenges remains to be elucidated. Mandarin Chinese, relying on pitch changes at the syllable level as lexical tones to convey meaning, is ideal for exploring this issue. Consequently, our study explores whether high autistic traits associate with poorer pitch-semantic integration among neurotypical Mandarin-speaking adults, employing the event-related potentials (ERPs) technique. Fifteen high-autistic-trait adults and eighteen low-autistic-trait counterparts participated in an implicit semantic priming task, and their behavioral and neural responses were recorded. Results showed that there existed significant differences in the reaction time between the two groups while no such differences were found for N400 responses, indicating similar processing mechanism despite different autistic traits. Our findings suggest that it might be other factors, such as verbal abilities, rather than autistic traits per se, that might influence the higher-level semantic integration related to pitch.

TAKEN BY SURPRISAL? ON THE ROLE OF LINGUISTIC PREDICTABILITY IN SPEECH RHYTHM

Tamara Rathcke, Chia-Yuan Lin, Eline Smit and Diego Frassinelli

Predictive processes and speech rhythm appear tightly interconnected during spoken language comprehension. Speech rhythm is often assumed to facilitate comprehension by shaping the time frame that influences neural entrainment and allows for temporal predictions to be formed. The present study asked the question about the relationship between linguistic predictability and rhythmic anticipation by using a rhythmic synchronization task. The task required participants to synchronize with the beat of spoken sentences repeated on a loop. The sentences differed in two sources of linguistic predictability, local (measured as surprisal of words within a sentence) and distal (implemented as repeating semantico-syntactic templates within a block of sentences). Thirty-two native speakers of British English took part in the study, tapping to the beat of sentences spoken in their native language. Signed asynchronies were measured from their synchronizations with nucleus onsets, to examine the degree of rhythmic anticipation. The results showed a greater degree of anticipation for the locally most informative parts of utterances within the distally most predictable context and could not be explained by bottom-up variability in acoustic salience (F0, duration, intensity). These findings suggest that the domain of predictively operating rhythmic attention in speech is linguistic rather than acoustic in nature.

The effects of regional Italian prosodic variation on modality identification by L1 English learners

Valentina De Iacovo and Paolo Mairano

Yes-no questions in Italian are not marked morpho-syntactically and intonation is the only cue distinguishing declarative vs interrogative modality. However, in different regional varieties of Italian, the intonation patterns of questions vary dramatically and yes-no questions can be realised with final rising or falling

contours. We investigate whether adult learners of L2 Italian correctly identify the modality (interrogative vs declarative) of a sentence, when pronounced by native speakers of different regional provenance, using different rising and falling contours. We developed an identification test where participants were exposed to 100 stimuli (10 sentences x 5 varieties x 2 modalities), pronounced by 10 speakers from 5 different regions in Italy. 20 L1 English learners of L2 Italian and 20 L1 Italian control speakers listened to the final syllables of each utterance and identified it as declarative or interrogative. Results show that L1 Italian speakers correctly identify sentence modality at higher rates than learners, and that questions with final falling contours have the lowest correct identification rates for learners. We argue that this may be attributed to L1 transfer (since a rise is the default realisation for yes-no questions in English, even more so without syntactic inversion), as well as to universal patterns.

Vocal and visual features in speech imitation

Sandra Madureira and Mario Fontes

Professional imitators incorporate in their speech lexical expressions and gramatical markers used by the speakers they imitate. They also change their voices and body movements to make them more similar to the imitated speakers and enable listeners to identify them. Voice quality and facial expressions are central aspects in speech imitation due to their indexical role. The objective of this paper is to investigate the changes in vocal quality and facial expressions introduced by a professional imitator to imitate public figures. The corpus comprises video excerpts of speech samples from a professional communicator and the public figures he imitates. The research methodology comprises: perceptual analysis of the vocal quality settings and prosodic features by means of the VPA systtem; automatic extraction of acoustic parameters with the Prosody Descriptor Extractor; automated analysis of the imitator and the imitated speakers‘facial expression Action Unities, and multivariate statistical analysis. Results show the changes in voice quality settings and Actions Unities introduced by the imitator to make his voice more similiar to that of the imitated speaker. The imitator´s mimic ability to make changes in his habitual voice quality and facial expression characteristics provides deictic input for listeners to identify the imitated speakers.

VISUAL CUES OF EMOTION EXPRESSION: PERCEPTUAL EVALUATION AND AUTOMATED SYSTEM ANALYSIS

Mario Augusto Souza Fontes, Sandra Madureira and Juliana Andreassa

The communicative relevance of visual cues has been highlighted in bimodal studies on speech perception of affective states and pragmatic meanings. These studies usually combine manual analysis of visual data based on FACS Action Unities and subjective perception tests. Automated analysis of facial expressions is also based on FACS and associates facial movements with affective expressions. The objective of this paper is analyzing facial expressions of four basic emotions by comparing the results of a perceptual test with those of an automated analysis based on the same stimuli. The research corpus comprises thirty utterances that were presented as video stimuli to thirty-four judges in a perceptual test as well as analyzed by means of an automated facial expression recognition software. The results of both evaluations are multidimensionally analyzed and theoretically discussed in relation to visual cue salience, valence and arousal features, and associations of Action Unities with affective states.

The Prosody of Polar vs. Alternative Questions in Urdu

Benazir Mumtaz and Miriam Butt

"This paper investigates the prosody of polar questions (PolQs) in comparison to alternative questions (AltQs) in Urdu. Inspired by \cite{bhattDayal2020}, who claim that Urdu/Hindi AltQs are disjunctions of PolQs from a semantic perspective, we examined this hypothesis from a prosodic perspective via three experiments. Our results clearly show that AltQs are not disjunctions of PolQs from a prosodic perspective. Our findings also resolve an existing disagreement with respect to the boundary tone of PolQs in Urdu/Hindi, indicating PolQs predominantly end with H\%, while AltQs end with L\%. More excitingly, we discovered that in addition to the three signature properties of AltQs already established cross-linguistically, our data indicate a fourth cue: the hat pattern.

With respect to the prosody-meaning interface, we also observed an interaction between question type and word class, with verbs showing longer duration in PolQs but NPs exhibiting longer duration in AltQs. We argue that this is motivated by the focus properties of the respective clauses. Finally, our judgment experiment with ambiguous stimuli challenges the assumption of a preference for PolQ interpretation, showing that an AltQ interpretation is more likely, particularly when case markers cliticize to the individual disjunctions, marking these as separate prosodic phrases."

Differential effects of word frequency and utterance position on the duration of tense and lax vowels in German

Ivan Yuen, Bistra Andreeva, Omnia Ibrahim and Bernd Moebius

Acoustic duration is subject to modification from multiple sources, for example, utterance position [13] and predictability such as occurrence frequency at word and syllable levels [e.g., 1, 2, 3]. A study of German radio corpus data showed that these two sources interact to modify syllable duration. Other studies have found that the predictability effect can percolate downstream to the segmental level, and that this downstream effect is sensitive to phonemic contrasts [8]. However, [5] showed that utterance-final lengthening is uniformly applied to tense and lax vowels in German. This then raises some questions as to whether the effects of the two sources of durational variation are uniformly applied or sensitive to phonemic identity.

The current study focused on the duration of tense and lax vowels in the stressed syllable of monosyllabic and disyllabic words in utterance-medial and utterance-final positions. Twenty German speakers participated in a question-answer elicitation task. A preliminary analysis of seven speakers showed effects of utterance position and word frequency, as well as interactions with vowel type, suggesting a non-uniform application of durational adjustments contingent on phonemic vowel identity. Interestingly, the frequency effect affects the duration of lax vowels, but utterance position affects the duration of tense vowels

Cues of voicing contrast in two Chinese dialects: Implication for sound change

Menghui Shi and Yiya Chen

In the literature on tonogenesis, it is commonly believed that as onset voicing-induced fundamental frequency (f0) differences are exaggerated, lexical tones would come into being, and the onset voicing distinction disappears. In some East and Southeast Asian tone languages, however, intermediate stages with tonal contrasts and voicing contrast co-existing have been reported. This study reports data from two Sinitic varieties (i.e., Shuangfeng Xiang and Lili Wu), which show the coexistence of the two types of laryngeal contrast, to shed light on what could be a possible diachronic pathway of changes in consonant voicing contrasts with lexical tones co-present in the system. Both varieties have obstruent voicing contrast and lexical tones; phonation serves to cue the voicing contrast. We found onsets from the voiced category consistently align with lower f0 contours across dialects and generations. However, the relationship between laryngeal timing (in terms of voice onset time) and phonatory state (in terms of contact quotient of vocal folds) varies, which suggests different patterns of cue reweighting for the phonetic implementation of voicing contrast, possibly reflecting different stages of how the voicing contrast may maintain or disappear.

Gestures time to vowel onset and change the acoustics of the word in Mandarin

Patrick Louis Rohrer, Yitian Hong and Hans Rutger Bosker

Recent research on multimodal language production has revealed that prominence in speech and gesture go hand-in-hand. Specifically, peaks in gesture (i.e., the apex) seem to closely coordinate with peaks in fundamental frequency (F0). The nature of this relationship may also be bi-directional, as it has also been shown that the production of gesture directly affects speech acoustics. However, most studies on the topic have largely focused on stress-based languages, where fundamental frequency has a prominence-lending function. Less work has been carried out on lexical tone languages such as Mandarin, where F0 is lexically distinctive.

In this study, four native Mandarin speakers were asked to produce single monosyllabic CV words, taken from minimal lexical tone triplets (e.g., /pi1/, /pi2/, /pi3/), either with or without a beat gesture. Our analyses of the timing of the gestures showed that the gesture apex most stably occurred near vowel onset, with consonantal duration being the strongest predictor of apex placement. Acoustic analyses revealed that words produced with gesture showed raised F0 contours, greater intensity, and shorter durations. These findings further our understanding of gesture-speech alignment in typologically diverse languages, and add to the discussion about multimodal prominence.

Acoustic-prosodic Analysis for Mandarin Disyllabic Words Conveying Vocal Emotions

Xuyi Wang and Hongwei Ding

This study conducted a comprehensive analysis of features using a validated audiometry corpus comprising 450 Mandarin Chinese disyllabic words across five emotional states: ``Angry," ``Sad," ``Happy," ``Fearful," and ``Neutral," produced by both male and female speakers. Employing machine-learning tools, the research identified and elucidated crucial acoustic-prosodic features for emotional vocalization. Results revealed several key points: First, the models showed that fear was acoustically the most recognizable emotion, while joy presented most difficulties. Second, in the identification of Mandarin emotional prosody, the spectrum characteristics like formant energy ratios were of primary significance, followed by those F0-related parameters such as the 20th and 80th percentiles of F0. Third, data of formant energy ratios mainly indicated that fearful voices were more turbulent, and those of F0-related features suggested a general increase in pitch for emotional speech. Moreover, considerable cross-speaker variations in affective vocalization strategies were observed, reflected in distinct feature patterns that our speakers exploited for their emotional expressions. Despite the considerable audio samples gathered from each speaker, the current corpus remains limited by its two-speaker scale. Nonetheless, ongoing efforts involve expanding the corpus with additional speakers. The scalability and replicability of the paradigm can facilitate seamless transplantation for future investigations.

An exploratory investigation of phonological and phonetic length contrasts perception in Italian vowels and consonants

Francesco Burroni and Pia Greca

Standard Italian is canonically described as a language that displays a phonological length contrast in the consonant system, but no corresponding contrast in the vowel system. Despite this fact, it is widely accepted that Italian vowels are phonetically lengthened by speakers in open stressed syllables, especially penultimate ones. However, no studies on the perception of both vowel and consonant length have been conducted. A crucial question remains open: do Italian listeners perceive the durational cues underlying a hypothesized phonological length contrast (for consonants) and a hypothesized phonetic contrast (for lengthened vowels) differently? We investigated this question in an online AX perception experiment with over a hundred Italian listeners. Results from a Mixed Effect Logistic regression model and Machine Learning classification showed that Italian listeners displayed indistinguishable identification functions for both the phonological length contrast of consonants and the “putative” phonetic durational contrast of vowels, meaning that perceptual discrimination of segmental duration was similar for phonologically long and short consonants and for vowels that were “phonetically” lengthened (or shortened) in open penultimate syllables. These results suggest therefore that Italian listeners discriminate differences in duration similarly for both consonants and vowels, either as a cue to phonological length contrasts or stress or both.

Implicit learning of tone-segment connections by adults with and without tonal language backgrounds

Xinbing Luo and Brechtje Post

Previous research has shown that adults can implicitly learn segmental phonotactic constraints, but the implicit learnability of prosodic features like lexical stress and lexical tone remains unclear, especially given the variability in results among learners from different language backgrounds. To address this, our study examined the implicit learning of non-native tone-segment associations in adults, both with and without tonal language backgrounds, specifically Mandarin and English. Initially, participants were evaluated on their ability to acoustically and lexically distinguish tones. Subsequently, they were unconsciously exposed to tone-segment connections in pseudowords, followed by tests to assess their memorization and generalization skills. We also measured their awareness of the tonotactic rules through confidence ratings and structural knowledge attributions.

The findings indicated that participants, irrespective of their tonal background, could effectively memorize and generalize patterns associated with contour tones but struggled with level tones. Mandarin speakers demonstrated a higher awareness of the rules. These results highlight that the type of tone and prior tonal knowledge significantly influence the ease of learning and the nature of the knowledge acquired, providing valuable insights into the implicit learning mechanisms of word prosody

Functional and phonetic determinants of categorical perception in two varieties of Chinese

Yang Yang, Carlos Gussenhoven, Victoria Reshetnikova and Marco van de Ven

A categorical perception experiment involving an identification and a discrimination task was administered to two groups of speakers of Chinese (Cantonese and Zhumadian Mandarin), in order to establish to what extent the results would reflect the phonetic form or the functional status of monosyllabic pitch contours. Each group was presented with stimuli produced by speakers of their own variety. Generally, a difference between monosyllabic pitch contours can express an intonation contrast (question vs statement) or a lexical contrast. We hypothesized that intonation contrasts are perceived gradiently, while lexical contrasts are discrete.

The Zhumadian Mandarin results were in perfect agreement with our working hypothesis. Because the Zhumadian lexical tone contrasts are expressed through variation in pitch shape and the intonation contrast is expressed through variation in pitch height, the Cantonese group was recruited to test our hypothesis with a language that expresses a lexical tone difference through pitch height variation and the intonation contrast through a pitch shape difference. The Cantonese lexical tone contrasts were perceived discretely, but so was the intonation contrast. We conclude that only intonation contrasts that are expressed by means of pitch height variation have no phonological representation, while intonation contrasts expressed through pitch shape differences and all lexical contrasts do. These results confirm widespread assumptions about the tonal representations of (lexical and intonational) pitch contrasts in a way that earlier experimental findings had failed to do.

Lexical Tone Perception and Comprehension in Mandarin-Speaking Children with Autism Spectrum Disorder

Ting Wang and Mengzhu Xu

As a basic perceptual attribute of sound and a key information carrier in both music and language, pitch plays a crucial role in the perception of speech and music. Nevertheless, individuals with autism spectrum disorder (ASD) tend to exhibit atypical pitch perception. While previous studies have noted slightly enhanced ability in autistic individuals' overall performance in nonspeech pitch perception, the perception of speech pitch, which is crucial for encoding lexical differences in tone languages such as Mandarin, remains unclear. Additionally, multiple studies have suggested that individuals with ASD suffer from abnormal auditory attention, potentially impacting their speech pitch perception negatively. The present study aimed to investigate the tone perception in Mandarin-speaking children with autism and identify whether attention shifting in individuals with ASD may exert an influence on tone perception. By means of two picture-matching tasks, the experiment assessed the lexical tone perception and comprehension of Mandarin-speaking children with ASD and typically developing (TD) controls, matched on age and nonverbal IQ. The results indicate a clear distinction between tone perception of children with ASD and TD controls. Besides, because of different patterns of attentional focus, difference in task types has certain effects on tone perception.

Hierarchical Intonation Modelling for Speech Synthesis using Legendre Polynomial Coefficients

Johannah O'Mahony, Niamh Corkey, Catherine Lai, Esther Klabbers and Simon King

Synthetic speech quality is now close to parity with human speech for isolated read speech utterances.There has therefore been a resurgence of interest in using speech synthesis for speech science research. However, many speech synthesis models lack control over prosody. The few models that are controllable do not use interpretable control values or controls that relate to prosodic theory. We present a model that enables control, by conditioning on a hierarchical Legendre polynomial representation of F0 at the phrase and word levels. The polynomial coefficients are data-driven but linguistically-motivated and have been used in previous studies of pitch accents and phrase contours. The coefficients are interpretable in their characterisation of the F0 contour because they describe mean F0, slope, and convexity. We demonstrate sufficient control of F0 to produce speech that is intonationally similar to a reference sample. Objective and subjective evaluations are used to compare our Legendre-conditioned model to a baseline, to a model conditioned on categorical prosodic features, and to an oracle model conditioned on ground-truth F0. Our model has lower F0 prediction error and higher correlation with ground-truth. Future work aims to apply these features to conversational speech, by learning Legendre coefficients from large speech corpora.

Predictive Modelling of perceptual strategies: exploring the perception of ironic tone of voice by L2 learners of French

Ziqi Zhou, Jalal Al-Tamimi and Hiyon Yoo

The use of acoustic correlates in the production of ironic tone of voice has been well-documented. However, how L2 learners employ these acoustic cues to decode ironic speech has been comparatively underexplored. This study aims to investigate the perceptual strategies utilized by native Mandarin speakers with advanced French proficiency to interpret the ironic tone of voice in French. 42 native Mandarin speakers participated in an irony identification task, during which they listened to utterances from a separate production task, designed to elicit ironic and non-ironic utterances. A predictive modelling approach was employed. Firstly, the results were subjected to Generalized Linear Mixed Models (GLMM), from which we calculated the Irony Score (I-score) to estimate the predicted probability that a specific utterance would be perceived as ironic. Subsequently, through a random forest regression analysis, we explored the relationship between the calculated I-scores and eight acoustic cues, recognized as essential correlates for ironic speech, as suggested by previous literature. Our findings suggested that F0 span is the most salient cue for native Mandarin speakers learning French as L2 in perceiving irony in French. In addition, jitter, speech rate, and intensity span carried relatively more weight than other acoustic cues for irony detection.

Stylised sustained prosody in three Australian languages

Kathleen Jepson, Rasmus Puggaard-Rode and John Mansfield

A similar, striking prosodic pattern is reported to occur in languages around Australia. It is characterised by a stretch of level, high pitch, and lengthening of the IP-final vowel. This pattern appears to have a similar meaning in each language, expressing the extension of an event in space or time. However, there are some differences in the form described. In this paper, we present a study of the acoustics of this pattern in three unrelated Australian Indigenous languages, and propose a method for automatically identifying examples within an audio file. Hence, the purpose of this paper is twofold: 1) to provide a cross-linguistic description of this prosodic pattern with the aim of acoustically describing cross-linguistic variation, and 2) to provide a proof-of-concept for a method to automatically identify this pattern which could allow other language data to be incorporated into the typological description in the future.

A stochastic dynamical system for pitch accents and its inversion

Khalil Iskarous, Jennifer Cole and Jeremy Steffman

The literature on the pitch accents of American English (AE) reveals substantial variation across speakers and within accent categories, as well as variation in which pitch accent category is produced in a given discourse context. In this work we present a stochastic revision of a deterministic dynamical system theory of American English pitch accents. This theory generates F0 trajectories from a system of differential equations that govern the change in F0 over time, capturing the distinctions in peak alignment and scaling that characterize within- and across-category variation in AE pitch accents. The stochastic model has one free parameter which is set by the language's phonological system. We also present a stochastic model of perception of pitch accents, which invokes the production model to generate hypotheses about the phonological free parameter describing the observed trajectory. We therefore aim to provide a framework in which variability can be explicitly modeled, and in which the interaction of phonology, production, and perception of prosody can also be modeled.

Exploring the Dynamics of Post Focus Compression in Bilingual Speakers: Evidence from Mandarin-Yangzhou Speakers

Zhihong Chen, Haoyun Chen and Wenxi Fei

Post Focus Compression (PFC) has been recognized as a potential typological tool for classifying languages into different language families. While its function has been well-documented in monolingual speakers, the dynamics of PFC in bilingual speakers remain less explored particularly when two languages with PFC (+PFC) are in contact. This study aims to fill this gap by investigating PFC in bilingual speakers of Standard Mandarin (MA) and Yangzhou Mandarin (YZ), two +PFC languages. Twelve bilingual Mandarin-Yangzhou speakers were recruited for a language background questionnaire and an experiment of producing both MA and YZ using a question-answer paradigm. Their F0, duration, and intensity were recorded for analysis. We found that proficiency in YZ was positively correlated with the presence of PFC in YZ and the loss of PFC in MA, while proficiency in MA showed the opposite trend. These findings indicate that when two +PFC languages interact in the same context, PFC is highly sensitive to the speakers' proficiency in each language, even for bilinguals. This study thus contributes to our understanding of the dynamics of PFC in bilingual speakers and its future applications in linguistic typology.

Perception of the merging tones in Taiyuan Jin Chinese

Zhenyi Liao and Lei Liang

Traditionally, there are five lexical tones in Taiyuan dialect, but previous studies have shown that some speakers merged the four falling tones T1/T2/T4/T5 in the production. The current study investigates how 25 participants (14 females) with different age groups perceive four falling tones by means of discrimination and identification tasks. Both accuracy rate and reaction time are measured. Results show that: 1) the stabilization sequence in perception is as follows: T2>T1>T4>T5, indicating that T2 (Shang sheng) is the most stable tone while T5 (Yang entering tone) is most susceptible to merge with other tones. 2) the T4 and T5 (Yin and Yang entering tones) exhibit the highest degree of merger, with a tendency to perceive T5 as T4. 3) while participants across three age groups demonstrate similar patterns of tone merger in perception, the extent of merger varies. In the discrimination task, the middle-aged and elderly groups exhibit a greater degree of T4-T5 merger compared to the young-aged group; in the identification task, middle-aged participants demonstrate the highest identification rate for non-entering tones but the lowest rate for the identification of the Yang entering tone, indicating a pronounced confusion between the Yin and Yang entering tones in perception among this group.

The role of prosody in pragmatic interpretation in Mandarin

Mengzhu Yan, Sasha Calhoun and Qi Tan

It has been well established that prosody plays a key role in sentence processing, influencing various aspects such as lexical activation, syntactic parsing, information structure marking, and the signalling of pragmatic information, including speech attitudes, acts, and emotions. In the current study, we particularly focus on whether accentuating the verb or noun in the construction “它看起来像X” (It looks like an X) yields different pragmatic meanings in Mandarin. In a rating experiment, Mandarin-speaking participants assessed the likelihood of the statement “It looks like an X” (e.g., It looks like a pencil) referring to a target (e.g., a pencil) in a picture that participants could not see. Results have shown a higher likelihood when the target noun carries a contrastive accent, as opposed to when the verb carries a contrastive accent, aligning with evidence from other languages (e.g., English). This study extends the existing research to Mandarin and offers novel evidence on Mandarin listeners’ categorization of intonation contours, indicating a functionally equivalent resource of “prosodic prominence” across languages, despite variations in the phonetic implementation (i.e. pitch accenting in English vs. pitch range expansion in Mandarin).

How does human hearing estimates sleepiness from speech?

Vincent P. Martin, Salin Nathan, Beaumard Colleen and Jean-Luc Rouas

Excessive sleepiness is a major public and personal health burden that would benefit from being measured in ecological and passive setups. Speech recording is implemented in all smartphones and is thus a relevant tool to do so. To evaluate the feasibility of detecting sleepiness from speech by the human hearing, two previous perceptual studies on 90 samples from the SLEEP corpus have been conducted (Huckvale et al. 2020, Martin et al. 2023), which yielded contrasting results. A way to investigate the origin of this disagreement would have been to study on which speech characteristics the listeners have based their estimation. However, none of these studies have collected such information.
In this study, we identify these characteristics by extracting speech features from the recordings, and training simple and explainable machine learning models to reproduce the annotation of each listener. Then, we measure the contribution of each feature to the decision of each model, and identify the most important ones. We then perform hierarchical clustering to draw listeners' profiles, depending on the features they rely on to identify sleepiness.

Investigating Tempo and Pause with Synchronous Speech

Paula Laine and Michael O'Dell

"Tempo changes typically affect pause durations more than articulation rate. We investigated this phenomenon using synchronous speech, a laboratory version of joint speech, which occurs naturally for example in demonstrations, sports events and religious gatherings.

Subjects synchronized with stimuli of varying tempo modified from original recordings by uniform stretching. Responses were compared to stimuli in terms of the number of pauses, durations of pauses and interpause intervals (IPI), as well as latencies measured at the beginning and end of the IPI.

All subjects successfully synchronized with all stimuli. Contrary to stimuli, responses contained more pauses at slower tempos, and the proportion of total pause duration was greater at slower tempos. Latencies at IPI end were greater for faster tempos and more negative for slower tempos. Latencies varied more at pause end than at IPI end.

Results support the assumption that pause is more elastic than articulated speech. Speakers tend to follow this pattern even when they successfully synchronize with stimuli created artificially with uniform stretching. Although speakers can produce pauses and IPIs which match the durations of the stimuli, they have difficulty when they need to combine pause and IPI durations appropriate to different tempos."

L2 Prosody Assessment by Combining Acoustic and Neural Model Features

Wenwei Dong, Roeland van Hout, Catia Cucchiarini and Helmer Strik

Computer-Assisted Language Learning systems are becoming increasingly popular, but most of the systems focus on the segmental level, while research on second language (L2) intelligibility emphasizes the important role of prosody. In this paper, we investigated possible methods to calculate L2 prosody scores automatically, using speechocean762, an L2 English corpus evaluated by experts at utterance-level for prosody, pronunciation accuracy and fluency. To develop an automatic L2 prosody assessment method, we first extracted 107 acoustic features, then applied regression analyses, followed by Lasso regression and Recursive Feature Elimination to select the most relevant features for prosody, fluency, and accuracy. We also explored a Kaldi-based acoustic model trained on native data to estimate L2 performance at the utterance level. The results showed that the combination of selected acoustic features and transformed Kaldi-based scores works best to predict the experts' evaluations. Prosodic features (loudness, duration, F0) are important for the prosodic evaluation, but also for fluency and accuracy. Other features play a role as well. Our outcomes show that L2 prosody is an important characteristic of L2 speech and that automatically obtained prosodic measures can be helpful in evaluating L2 performance.

Does communicative skill predict individual variability in the prosodic encoding of lexical and referential givenness?

Janne Lorenzen and Stefan Baumann

We investigated individual variability in the prosodic encoding of lexical and referential givenness in German. Additionally, we related this variability to self-assessed communicative skill in our participants. In an interactive reading task, we collected data from 20 speakers producing eight short stories. In each story, the same target word occurred as lexically (l-) and referentially (r-)new or given, and in combinations of these levels. We measured several prosodic correlates of prominence in the target words. Across speakers, l-new referents were marked by longer duration and higher periodic energy than l-given referents, while r-new referents were marked by higher periodic energy and higher intensity than r-given ones. While speakers were remarkably similar in their encoding of l-givenness, only differing in how strongly they modified duration and periodic energy, there was a more striking contrast in the encoding of r-givenness: One group of speakers exclusively relied on periodic energy and intensity, the other group additionally used higher F0 to mark r-newness. Differences in the produced givenness contrasts across speakers proved to be related to communicative skill, albeit in opposite ways: L-givenness was marked more strongly by speakers with higher communicative skill, r-givenness was marked more strongly by speakers with lower communicative skill.

Is there an uncanny valley for speech? Investigating listeners' evaluations of realistic TTS voices

Alice Ross, Martin Corley and Catherine Lai

The exploration of uncanny valley effects (UVE) - a distaste for entities that appear almost, but not quite, human - has been a productive topic of research in human-robot interaction. Meanwhile, realistic text-to-speech (TTS) voices are increasingly encountered in various settings. In this work, we aim to describe the relationship between the perceived human-likeness and pleasantness of TTS voices and seek evidence of auditory UVE in listeners' evaluations.

In an online between-subjects experiment, listeners rated an array of manipulated TTS voices, trained using a single speaker's data. The evidence obtained is compatible with a slight plateau in a generally positive correlation between realism and approval. All the TTS voices used received ratings of below 50% on average for 'human-likeness', and therefore conclusions about UVE, i.e. negative reactions to voices perceived as very human-like, cannot be drawn from these data. Our results suggest that, although a correlation exists, high realism may not be necessary for relatively high approval; on average, voices with decreased pitch variation were rated about twice as highly for being 'pleasant' and 'friendly' as they were 'like a human'. The relationship between pitch variation and perceived realism is examined and identified as a direction for further research.

Emergent Dialectal Patterns: Analysis of regional variants in a vast corpus of Finnish spontaneous speech using a large-scale self-supervised model

Tuukka Törö, Antti Suni and Juraj Šimko

Traditional linguistic analyses, focused on morphological, syntactic and lexical features, as well as phoneme-level differences, divide Finnish into two major dialect groups, that subsequently further split into eight sub-groups. This paper presents a complementary dialectal analysis based solely on acoustic characteristics extracted from an extensive database of spontaneous speech from thousands of speakers from all Finnish dialectal areas. The distances among acoustic characteristics of speech from 17 administrative regions are approximated by prediction accuracies of binary classifiers. The classifiers are trained on principal components extracted from utterance embeddings obtained through a large-scale pretrained neural model. The clustering of regional varieties based on these distances yields geographically meaningful dialectal groupings, largely corresponding to the results of the traditional linguistic analyses. Our subsequent analysis indicates that the clustering makes use of prosodic characteristics of utterances.

Prosodic realization of rhetorical and information-seeking questions in Persian

Mortaza Taheri-Ardali, Tina Ghaemi, Tina Bögel and Bettina Braun

This paper investigates the prosodic realization of rhetorical questions (RQs) in comparison to information-seeking questions (ISQs) in Persian, an Iranian language. In a lab setting, we recorded polar questions (word order: S-O-V) and wh-questions (word order: wh-O-V) in information-seeking and rhetorical contexts. We then analyzed constituent durations, voice quality (breathy, modal, glottalized), and the intonational realization. The results showed that all constituents, except the verb in wh-questions, were significantly lengthened in RQs compared to ISQs. Furthermore, RQs were more often realized with breathy voice quality than ISQs. Intonationally, RQs had lower f0-values than ISQs, in particular towards the end of the questions. This was caused by phonological differences: Polar questions, which were primarily realized with high or downstepped high boundary tones (H%, !H%), differed significantly in whether there was an accent on the verb (ISQ) or not (RQ). In wh-questions, in which the object and the verb were typically deaccented, differed in boundary tone: wh-RQs most often had L% boundary tones and wh-ISQs !H%. The results for Persian align with findings for other, typologically different, languages. Furthermore, they provide data that have not been discussed for Persian yet (e.g., !H% in wh-questions).

Use of lexical stress information in German learners of English

Alexandra Jesse, Julia Sigg and Ulrike Domahs

German and English are both lexical stress languages, but differ in the relative use of suprasegmental and segmental cues. The languages also tend to highlight different syllables within polysyllabic words. While Romance borrowings in German exhibit predominantly primary stress at the right edge of words (e.g., Admirál), analogous borrowings in English are very often stressed on the initial syllable (e.g., ádmiral). The question arises whether differences in primary stress position have an impact on L2 word processing. In the present study, German advanced learners of English were tested on their recognition of polysyllabic English words in a visual-world paradigm. Eye fixations were recorded while hearing instructions to click on one of four written words presented on a screen. Critical pairs overlapped segmentally for at least two syllables and had either primary stress on the initial syllable (ádmiral) or on the third syllable, with initial secondary stress (àdmirátion). German listeners used primary stress, but not secondary stress, for online word recognition. They also showed, unlike English listeners, a bias before target word onset towards words with stress in later syllables, potentially reflecting L1 metrical expectations, hence whether German listeners can use primary stress to facilitate early recognition remains unclear.

Prosodic marking of contrastive focus in French learners of German

Katharina Zahner-Ritter, Nathalie Elsässer, Ingo Feldhausen and Jürgen Trouvain

Challenges in the foreign language acquisition of intonation often result from cross-linguistic differences. While German, for example, encodes contrastive focus via intonation (pitch accents), French typically employs a syntactic strategy (clefting). For these reasons, the present study investigates the intonational marking of contrastive focus in German using learners with French as their native language via recordings of read sentences from the IFCASL corpus. Productions of learners (34 beginners and 34 advanced learners) were prosodically analyzed (placement, type, and prominence of the accent), and compared to productions of 40 native speakers of German. Results revealed that though pitch accent types differed from those used by native speakers, learners generally succeeded in placing pitch accents on the contrastively focused element. However, by far the greatest difficulty for learners at all levels concerned the deaccentuation of non-focal elements. In this respect, advanced learners were more successful in reducing prosodic prominence in the pre- and post-focal regions. Our findings are crucial for developing tailored teaching materials that concentrate on non-focused elements to enhance learners' mastery of target-like prosodic marking of contrastive focus.

Does Tone Impact Mandarin Non-Word Acceptability Judgements?

Jamie Adams, Sam Hellmuth and Leah Roberts

Using non-words in psycholinguistic research allows for a high level of control over experimental stimuli. However, this relies on the assumption that they reflect natural language. Eliciting acceptability judgements from L1 speakers of the target language is one approach to ensuring the relative authenticity of stimuli. For tonal languages, it is as yet unclear whether tone interacts with the perceived acceptability of non-words. In this between-participant Mandarin non-word norming study, 72 L1 Mandarin listeners judged 750 syllables across five tones: tones 1-4 and the neutral tone (NT). Syllables were analysed as systematic gaps, which do not appear in the lexicon because they violate phonotactic constraints, and accidental gaps, which are phonotactically sound but are absent from the lexicon. Real words and malformed syllables acted as maximally and minimally acceptable controls, respectively. Linear mixed effects models indicate that tones 1-4 do not modulate acceptability judgements. NT had a significant negative effect, but this likely arises from exposure to excised neutral tone syllables out of context rather than ungrammaticality. We suggest that Mandarin non-words can be associated with any lexical tone without concern for its effect on acceptability but that neutral tone stimuli should be presented in context to preserve authenticity.

Native and Non-native listeners' Ability in Integrating Prosody and Verb Semantics in Mandarin Speech Comprehension under the Impact of Language-specific Prosodic System

Xiaomu Ren and Clara Cohen

This visual-world eye-tracking study examined the prosody and verb semantics integration in native and non-native Mandarin speech perception. Native Mandarin and English listeners' eye movements were recorded and they were asked to click on the object named in the second sentence while they listened to Mandarin sentences within a brief discourse context. These sentences varied in three binary factors: prosodic accent, target object information status, and semantic match between verb and target object. Results showed a complex interaction between L1, prosody, and verb semantics. When processing both old and new information, Mandarin listeners interactively combined high-level verb semantic cues with low-level prosodic cues across diverse contexts, such that effects of verb semantics were exaggerated under contrastive focus. English listeners showed a similar pattern with old information as Mandarin listeners, but with new information they showed less integration of different speech cues than Mandarin listeners. These results suggest that non-native listeners can adopt native-like strategies in integrating prosodic and semantic speech cues, but only when the information status of target words is familiar, and hence less tasking to process.

Investigating lexical stress accuracy in non-native speech through real-time speech visualization: a pilot study

Kizzi Edensor Costille

This pilot study explores the impact of visualizing real-time speech on improving lexical stress among non-native English speakers. The study introduces a real-time 3D spectrogram, where learners can see, hear and imitate a model's speech and view their own productions. Six French English learners participated in a three-phase within-subject study consisting in a pre-test, a 10-week training session using the spectrogram, and a post-test. The study questions whether visualizing speech improves lexical stress production and equips learners to handle new words post-training. An auditory analysis of pre-test and post-test results revealed a slight improvement in correct lexical stress placement, with the global mean of accurately pronounced words rising from 4 in the pre-test to 4.5 participants in the post-test. The mean for correctly pronouncing pre-test words included in the post-test, improved by 1 (from 4 to 5) but there was minimal improvement in correctly pronouncing the new words in the post test. The goal of this study is to contribute to understanding how L2 learners can improve their word stress accuracy in English and to expand our knowledge regarding multi-sensory tools' efficacy in second language learning.

Investigating the causes of prosodic marking in self-repairs: an automatic process?

Morgane Peirolo, Candice Frances and Antje Meyer

Natural speech involves repair. These repairs are often highlighted through prosodic marking (Levelt & Cutler, 1983). Prosodic marking usually entails an increase in pitch, loudness, and/or duration that draws attention to the corrected word. While it is established that natural self-repairs typically elicit prosodic marking, the exact cause of this is unclear. This study investigates whether producing a prosodic marking emerges from an automatic correction process or has a communicative purpose. In the current study, we elicit corrections to test whether all self-corrections elicit prosodic marking. Participants carried out a picture-naming task in which they described two images presented on-screen. To prompt self-correction, the second image was altered in some cases, requiring participants to abandon their initial utterance and correct their description to match the new image. This manipulation was compared to a control condition in which only the orientation of the object would change, eliciting no self-correction while still presenting a visual change. We found that the replacement of the item did not elicit a prosodic marking, regardless of the type of change. Theoretical implications and research directions are discussed, in particular theories of prosodic planning.

Levelt, W. J. M., & Cutler, A. (1983). Prosodic Marking in Speech Repair.

Is automatic phoneme recognition suitable for speech analysis? Temporal and performance evaluation of an Automatic Speech Recognition model in spontaneous French

Vincent P. Martin, Colleen Beaumard, Jean-Luc Rouas and Yaru Wu

The correct automatic identification and segmentation of phonemes is crucial for a more in-depth exploration of prosodic parameters on a syllabic level. As such, automatic phonemic transcription from spontaneous speech recordings has numerous applications, such as teaching or health monitoring. Such transcriptions are usually evaluated either in terms of correct phoneme estimation or temporal segmentation, each task being addressed by a dedicated system. However, no system to our knowledge has ever been evaluated on doing correctly the two tasks at the same time. This article evaluates a state-of-the-art Kaldi-based phonetic transcription system for spontaneous French.

We use the Rhapsodie database, composed of spontaneous speech recordings with diverse levels of planning. Our phoneme recognition system obtains good results on phoneme and phoneme category identification (respective error rates of 19.2% and 13.4%), performed poorly on phonemes and category segmentation: an average of 40% of phoneme duration and 34% of phonetic categories duration have not been detected by it. On both metrics, the performances of the system increase with the degree of planning of the spontaneous speech.

These results suggest that improvements are necessary for designing truly reliable automatic phonetic transcription systems to be useful for further analysis.

Speech Adaptation and Physiological Responses: A Study on f0 and Skin Temperature

Tom Offrede, Christine Mooshammer, Alessandro D'Ausilio and Susanne Fuchs

This study investigates f0 adaptation, skin temperature change, and the relationship between the two. While a growing number of studies have demonstrated that emotional reactions in humans lead to changes in their facial skin temperature, none of them have studied temperature change in conversational contexts. Here, we have tested whether a conversation's degree of intimacy influences emotion such that it affects facial temperature and f0 adaptation—in terms of entrainment to interlocutor and f0 change due to the conversation topic. We also ask whether temperature change and f0 adaptation are related. In our data set of 38 participants in a between-subjects design, few speakers aligned on f0 to their partner, with no identifiable patterns. Regardless of their interlocutor, however, the speakers' f0 median and standard deviation tended to decrease when they spoke about more personal topics. This adds to previous literature describing emotional speech prosody. The participants' nose temperature was modulated by social emotion, but there was no relationship between temperature change and f0 adaptation. This suggests that, although the participants' nose temperature was sensitive to the social dynamic, the emotional reactions driving thermal change do not seem to be the same leading to prosodic adaptation.

Prosodic prominence and its hindering effect on word recall in German

Barbara Zeyer and Martina Penke

The aim of our study was to deepen our understanding of the influence of prosodic prominence on language processing. We conducted a word recognition memory task in German where we manipulated the word preceding the target word with either the highest (L+H*) or the lowest prominent accent type (deaccentuation) in German. Previous studies have shown higher accuracy rates when the target word was prosodically manipulated with a high prominent accent type. Based on these findings, we postulated that prosodic prominence binds processing resources, as the attention of the listener is drawn to the prosodically prominent entity, furthering recognition memory of this entity. Conversely, we assumed that a highly prominent accent (L+H*) on the word preceding the target word would lead to less accurate recognition rates of the target word compared to a control condition with deaccentuation. In this case, processing of the prominently accented word would bind attention and processing resources that would not be available for a deeper anchoring of the following word, the target, in memory. Our data confirmed this expectation. Prosodic prominence hindered recognition when it was on the word preceding the target word. This finding supports our assumption that prosodic prominence binds attention and processing resources.

The effect of age and gender on global intonational features in heritage and monolingual Russian

Yulia Zuban and Sabine Zerbian

The present study examines the effect of age and gender on global intonational features in a prosodically annotated corpus of spontaneous narrations by adult and adolescent heritage speakers (HSs) of Russian in the U.S. and monolingual speakers of Russian. The study investigates the use of boundary tones (BTs), bitonal vs. monotonal pitch accents (PAs) and frequency of PAs in these speaker groups.

The results show that age and gender influence some of the investigated intonational patterns, but not all. Specifically, it was found that female and adolescent speakers in general produce high BTs more frequently than male and adult speakers. Further, female HSs produce more monotonal accents than monolingual female speakers while male heritage and monolingual speakers are similar to each other. Also, monolingual female speakers produce fewer monotonal accents compared to male monolingual speakers. Finally, PA frequency was not found to be influenced by age or gender.

The results are discussed with reference to physiological differences between male and female speakers, possible influence from the HSs’ majority language English and possible ongoing change in monolingual Russian that is initiated by female and younger speakers.

Unsupervised modeling of vowel harmony using WaveGAN

Sneha Ray Barman, Shakuntala Mahanta and Neeraj Kumar Sharma

Neural network models of phonological learnability are said to learn the phonotactics of a language better than traditional models of learnability[1]. Our paper explores whether the Featural InfoWaveGAN architecture (fiwGAN [2]; inspired by WaveGAN [3] and InfoGAN [4]) can capture regressive vowel harmony patterns when trained unsupervised on raw acoustic data without any supply of prosodic cues. We train the model with Assamese speech data recorded by 15 native speakers. Assamese is one of the few Indian languages that exhibit phonologically regressive and word-bound vowel harmony. [+high, +ATR] vowels [i, u] trigger right-to-left harmony of [-ATR] vowels [ɛ, ɔ, ʊ] resulting in [e], [o], and [u], respectively. We analyze the outputs generated by the fiwGAN model and observe that it learns the regressive directionality of harmony. It produces innovative items by stringing together vowels and consonants from the training dataset. It showcases its capability of learning the phonotactics of Assamese and iterative harmony patterns over a longer domain without any relevant prosodic information in the output. We assume the model treats the outputs as abstract prosodic units without external prosodic cues triggering vowel harmony.

Acoustic classification of speech with trustworthy intent

Constantina Maltezou-Papastylianou, Reinhold Scherer and Silke Paulmann

Non-verbal prosodic patterns in speech have the power to communicate a speaker's emotional state, health condition, gender and even personality traits, such as trustworthiness. While research has mainly focused on the relationship between speech acoustics and perceived personality traits from a listener's perspective, the current research has developed a large speech dataset to examine the production of speech with the intent of sounding trustworthy, based on the speakers' self-perception. More precisely, the current research is looking to identify whether certain acoustic cues can be used to characterise a speaker's intent (i.e. neutral or trustworthy). In total, ninety-six younger and older adults from different ethnic backgrounds (i.e. white, black and south Asian) were recruited. They were asked to initially speak a set of sentences in their normal way of speaking ("neutral") and then repeated the same sentences, but this time they were asked to convey the intent of sounding trustworthy. Our findings provide evidence that pitch and voice quality related features can correctly differentiate a speaker's intent from our audio dataset with an accuracy of ~70%.

PitchMendR: A Tool for the Diagnosis and Treatment of F0 Irregularities

Thomas Sostarics and Jennifer Cole

Irregularities in F0 tracking such as sudden jumps or the halving/doubling of F0 often arise from consonantal perturbations, voice quality modulations, or environmental noise. These irregularities are typically visually apparent to the researcher, but fixing such errors is a time-intensive process even with algorithms that provide heuristic assessments of potential errors. In this paper we describe PitchMendR: an R-based interactive visualization tool to rapidly identify and fix irregularities. We discuss the main features of the tool and a proof-of-concept analysis of how it can be used to reduce noise in a dataset for statistical modeling.

Phonetic and phonological factors in cross-dialectal tone perception

Wenqi Zeng and Christine Shea

The perception of novel tones is influenced by previous experience with a tonal system. While cross-linguistic tonal acquisition has been examined in previous work, the situation of bidialectal tonal speakers has received less attention, and little is known about how two established tonal systems interact and shape the perceptual space.

We examined Chengdu Mandarin-Standard Mandarin (CM-SM) bidialectal and Standard Mandarin (SM) monodialectal speakers' perceptions of tones across dialects. Both CM and SM have four tone categories with cross-dialectal one-to-one phonemic correspondence. Importantly, each tone category has different pitch realizations between the two dialects. Therefore, in the case of CM-SM bidialectal speakers, both phonetic and phonological factors could potentially play into cross-dialect perception.

The results from a pair-wise dissimilarity judgment task show that for both groups, acoustic-phonetic similarity was the driving factor in dissimilarity judgments. However, the bidialectal group's judgment was also influenced by phonological factors. Compared to the monodialectal group, the bidialectal group perceived two tones with cross-dialectal phonemic correspondence as more similar to each other.

This study shows that bidialectal speakers categorize their tonal systems differently from monodialectal speakers, even when their two tonal systems differ primarily at the level of phonetics.

Perceptual salience of tonal speech errors

Zifeng Liu, Ioana Chitoran and Giuseppina Turco

The present study examines the perceptual salience of tonal speech errors compared to segmental errors (consonant & vowel). Tonal errors are observed less often than segmental errors. We thus hypothesize that tone errors are more easily ignored during transcription tasks because tones have lower perceptual salience relative to segments. We test this hypothesis in Mandarin, via a number reconstruction task. Sixty-nine Mandarin native listeners heard sequences of numbers in which one number was altered by substituting its vowel, consonant, or tone. They were asked to identify which number that was. We found that Mandarin listeners identified the original number most accurately when consonants were substituted. They were the least accurate when vowels were substituted. For tone substitution, the accuracy was lower than for consonant substitution, but not significantly different from vowel substitutions. Reaction times to identify a number with tone substitution were comparable to those for other types of substitutions. The results show that, contrary to our hypothesis, tone errors are not perceptually less salient than segmental errors. Specifically, tone errors are as salient as vowel errors and more salient than consonant errors, suggesting a similar phonological status shared by tone, vowel and consonant.

Neural tracking of prosodic structure in delexicalized speech

André Bernardo, Pedro Correia, Marina Vigário, Ricardo Vigário and Sónia Frota

The cortical oscillatory framework proposes that neural oscillations at different frequency bands track linguistic structure in speech. Some scholars argue that neural tracking can be dissociated to some extent from the encoding of cues in the speech signal, whereas others have emphasized neural tracking at different timescales, matching stimuli specific properties. The modulations of the amplitude envelope, which mostly capture prosodic information, have been shown to play a key role in the process, suggesting that prosody is central to the neural tracking of speech. The aim of this study was to isolate prosodic processing from morphosyntactic and semantic processing to investigate neural tracking of prosodic structure. Using delexicalized speech (with the original syllables replaced by a single syllable, e.g., /mi/), we show that neurophysiological signals in the delta and theta bands track the rates of the utterance (~0.4-0.7Hz), the intonational phrase (IP, ~0.9-1.2Hz), the prosodic word/stress (~1.5-2.5Hz), and the syllable (~4-6Hz). Moreover, the evoked power spectrum of EEG responses distinguished between stimuli with a one-IP-utterance (larger power at the utterance rate) and stimuli with the utterance comprising two IPs (larger power at the IP rate). These results indicate that prosody alone is sufficient to drive neural tracking at different timescales.

Test-retest reliability of audiovisual lexical stress perception after >1.5 years

Floris Cos, Ronny Bujok and Hans Rutger Bosker

In natural communication, we typically both see and hear our conversation partner. Speech comprehension thus requires the integration of auditory and visual information from the speech signal. This is for instance evidenced by the Manual McGurk effect, where the perception of lexical stress is biased towards the syllable that has a beat gesture aligned to it. However, there is considerable individual variation in how heavily gestural timing is weighed as a cue to stress. To assess within-individual consistency, this study investigated the test-retest reliability of the Manual McGurk effect. We reran an earlier Manual McGurk experiment with the same participants, over 1.5 years later. At the group level, we successfully replicated the Manual McGurk effect with a similar effect size. However, a correlation of the by-participant effect sizes in the two identical experiments indicated that there was only a weak correlation between both tests, suggesting that the weighing of gestural information in the perception of lexical stress is stable at the group level, but less so in individuals. Findings are discussed in comparison to other measures of audiovisual integration in speech perception.

Melodies of learning: A prosodic analysis of preschool teachers' language patterns in the classroom

Jill Thorson and Kim Nesbitt

Rich linguistic environments (RLEs) are essential to support child and language development in early education settings. Characteristics of RLEs include high-quality language, prosody (fundamental frequency-f0 variation), and interactions between the teacher and children (increases in turn-taking, number of contingent responses, and open-ended questions). We currently lack a quantifiable way to assess teacher language to provide clearer feedback to enhance early learning environments. The aims are threefold: 1) to quantitatively assess the quality of teacher language, 2) to analyze prosodic variation across a typical day for a preschool teacher, and 3) to investigate connections between teacher language and prosody. Audio data were recorded from preschool classroom teachers from 5 different sites (preschool teachers of 3-4-year-old children; n=10). Each teacher was recorded for two days (4-7 hours/day). Each participant completed a background survey and a daily activity log. The recordings are transcribed to analyze language (MLU, number of questions, number of different/total words, number of turns) and prosodic characteristics (f0 variation; prosodic contour shapes). Data from this study aid in understanding RLEs and provide a method to systematically capture teacher language in a more nuanced manner.

A prosodic approach of constructed action in Belgian French

Hadrien Cousin

Constructed action is a communicative strategy whereby speakers use their bodies and voices to depict referents, actions, and thoughts. By using a corpus-based study, this research analyses the use of constructed action and constructed dialogue in the production of four speakers of Belgian French. Previous research has rarely focused on the prosodic aspects of constructed action. Therefore, the aim of this study is to analyse constructed action with a special focus on prosodic components. The first goal is to understand how constructed action is prosodically marked. Prosodic elements can be assumed by the body (i.e. a head-nod that accompanies a stress). Indeed, information is not only exchanged via the auditory channel, but also via a wide range of visual cues. Therefore, the present study analyses as much as vocal than visual prosody. The results show that when speakers of Belgian French use constructed action to enact referents, they draw on a combination of multiple articulators. Speakers use different parts of their body (head, shoulders, and lips), linguistics elements (like verbs), but also fundamental frequency, intensity, and vocal quality. The results also suggest that speakers use more frequently their voice than other parts of the body to depict referents.

Mandarin Tonal Contours in Speakers with Autism Spectrum Disorders (ASD): Insights into Informational Structure

Vanessa Shih-Han Wu, Hohsien Pan and Susan Shur-Fen Gau

Previous research did not identify f0 as a cue distinguishing new and given information produced by native Mandarin speakers with Autism Spectrum Disorder (ASD). This study investigate the effect of information structures on f0 onsets and f0 ranges in new and given information bearing syllables from spontaneous dialogues elicited during the emotion session of Autism Diagnostic Observation Schedule (ADOS). The log semitone f0 at vowel onsets and f0 ranges were divided into four subsets according to four lexical tones. The results indicated that non-autistic and autistic speakers did not distinguish information structures with f0 cues. Instead, male autistic adults produced prominent new information-bearing syllables with significantly narrower f0 ranges and lower f0 onsets for dynamic rising tone T2 or falling tone T4 than TD speakers did. These findings align with the previous descriptiond which found that f0 cues in general do not mark prosodic porminent new information bearing syllables.

The prosody of theticity and focus in Beginner L2 Spanish

Sebastian Leal-Arenas and Marta Ortega-Llebaria

To express focus on the subject, English speakers enhance the duration, F0, and intensity of the word-in-focus, e.g., “ANNE is sleeping”. In contrast, Spanish speakers use verb-subject inversion to express an equivalent meaning, e.g., “Duerme Ana”. Similarly, thetic-categorical meanings are expressed via intonation in English and word-order in Spanish. Even at advanced proficiency levels, English learners of Spanish continue transferring their L1 intonation strategy while struggling to use word-order to convey those meanings. The present study investigates whether explicit instruction to elementary Spanish learners on the use of word inversion to express subject-verb focus and thetic-categorical contrasts leads to the suppression of English in-situ prominences, and consequently, to the adoption of Spanish intonation. Thirty learners, divided into experimental (N = 20) and control (N = 10) groups, completed a pre-test, a 10-session training, and a post-test. Participants were asked to produce complete sentences with subjects and verbs. Results revealed that word-order accuracy improved in the experimental group. Acoustic intonation analysis showed that only those participants with accurate word-order and no pauses stopped transferring L1 in-situ intonation approaching monolingual values. This suggests that the learning of word-order precedes the learning of Spanish sentence melody.

Word Stress and Prosodic Events in Eastern Armenian

Samuel Chakmakjian, Hossep Dolatian and Stavros Skopeteas

In languages with word-final stress, it is always a challenge to distinguish whether prosodic events are associated with the word-final stress (pitch accents) or the right edge of the prosodic word (edge tones). The present study focuses on Eastern Armenian, where stress occurs within the last non-schwa syllable – excluding certain unstressed suffixes. To determine the associate of nuclear/prenuclear tonal events, we conducted a speech production experiment (scripted speech) with 10 native speakers of Eastern Armenian. Target words, varying in stress placement (final, penult), were recorded as objects in SOV sentences. Utterances with these targets were performed as answers to questions that had different types of focus: pre-nuclear (V-focus) and nuclear (O-focus). Our Autosegmental-Metrical analysis, complemented by aggregated F0 contours, reveals that (a) a high pitch target demarcates the pre-nuclear domain from the nucleus, aligning with the right edge of the target word; (b) a high pitch target associates with the stressed syllable of the nucleus. These findings allow us to establish a distinction between two types of H-targets in Eastern Armenian: edge tones delimiting pre-nuclear material (H-) and pitch accents (H*) at the intonational nucleus.

Rhythm and the role of rhythmic variation in speech recognition: Analysis of African American English

Li-Fang Lai and Nicole Holliday

Although the past decade has witnessed rapid advancements in automatic speech recognition (ASR) technologies, how prosodic variation impacts errors and why systems return degraded performance for African American English (AAE) speech are still not well-understood. The present study conducted two sets of analysis to offer insights into these issues. First, we computed seven quantitative measures of rhythmic variation (%V, ΔC, ΔV, VarcoC, VarcoV, nPVI-V, and rPVI-C) in a reading task produced by AAE speakers and tested the effect of articulation rate on these metrics. The results reveal intricate interactions between articulation rate and prosodic rhythm as found in non-AAE speech, while at the same time showing timing properties specific to AAE speech. We then examined the seven metrics and their relationships to word error rates. Results show that utterances exhibiting shorter %V and greater VarcoV values had higher error rates. We argue that shorter %V and greater VarcoV can be explained through vowel reduction in unstressed vowels, repetition reduction effect, and monophthongal /aɪ/ and /ɔɪ/, a well-documented AAE feature that may contribute to recognition errors. The results suggest that adding rhythmic variation to ASR acoustic models can provide additional information for developers interested in mitigating racial bias in voice technology.

Prosodic Clitics in English-speaking Children's Speech Production – An Acoustic Study

Rui Cai, Paul Boersma, Ivan Yuen, Katherine Demuth and Titia Benders

English-speaking children have been suggested to cliticize function words as early as 2 years of age. However, several limitations to previous research, notably the absence of identical acoustic measures of cliticization across all ages, pose challenges in elucidating apparent differences between 2-year-olds and school-aged children. Thus, this study aims to apply established methods from adult research to children's speech and provide insight into how children acoustically realize cliticization in their productions. The study explored the production of cliticization by comparing two prosodic structures: No-Clitic (e.g., Boys often cut cards) vs. Potential-Clitic (e.g., Boys often cut the cards). A total of 32 children, 3-year-olds (N=12) and 12-year-olds (N=20), were drawn from AusKidTalk, an audio corpus of Australian-English children's speech. This selection of the corpus' youngest and oldest age groups enables an initial exploration of age-related change in cliticization. The results show that, on average across the ages, children shorten verb durations in the Potential-Clitic condition compared to the No-Clitic condition, indicating that children cliticize articles in the leftward direction. Our findings then further suggest that a trading relationship between verbs and articles exists in the process of children's acquisition.

Cluster analysis of Korean IP-final intonation

Hae-Sung Jeon, Constantijn Kaland and Martine Grice

In the Korean Tones and Break Indices (K-ToBI) system, Intonational Phrase (IP) final f0 contours are represented using tones at two levels, Low and High. However, functional analyses have shown the need for two additional levels, Mid and Top. These four levels are also used in some speech synthesis applications. This paper uses hierarchical clustering to investigate the f0 contours on IP final syllables in spontaneous dialogues. Our aim is to develop a classification scheme for analysing surface-level intonational variation in Korean. The clustering results show differences in register for the IP final f0 contours, supporting the usefulness of a phonetic analysis with more than two levels. The IP's position in a sentence seems to affect the IP boundary tone; the low boundary tone for the sentence-medial IP may not be as low as that for the sentence-final IP. The mapping between meaning and boundary tones is not straightforward, and the effect of post-positional particles must be considered when investigating intonational meanings in Korean. Given the variability related to morphosyntactic and sociolinguistic factors, implementing a fine-grained classification scheme that goes beyond two levels is desirable for further exploration of large corpora.

The effects of period doubling and vocal fry on the perceived naturalness of Mandarin tones

Yaqian Huang

Voice quality cues are used in tone perception. For Mandarin tones, previous studies have found that creak can facilitate or improve the identification of the low-dipping tone 3 [1,2]. Low f0 in particular was found one of the most useful cues of creak, and resynthesized period doubling was found to hinder all tone perception [2]. However, it is less clear how naturally-occurring subtypes of creak affect tone perception. Here we focus on the effects of period doubling and vocal fry, compared to modal voice, on the perceived naturalness and representativeness of Mandarin tones. A visual sort-and-rate task was implemented using naturalistic tones with varying voice qualities. Native Mandarin speakers rated tone 3 as more natural when having vocal fry or period doubling with frequency and amplitude modulation but other tones as more natural when having modal voice. The rising tone 2 and falling tone 4 were also rated lower when having period doubling with amplitude modulation. Interestingly, the high-level tone 1 and tone 3 with amplitude-modulated period doubling had similar ratings with modal voice. These results suggest that period doubling has versatile functions in different tone contours depending on the modulation type, and creaky voice is integrated in tone 3.

Question Intonation in Guanzhong Mandarin

Jiarui Zhang

This study examines the intonational tunes of syntactically unmarked polar questions in Guanzhong Mandarin (GuanM), particularly focusing on the exploration of the boundary tone in Chinese languages and the interaction between tone sandhi rules and intonation. We conducted a production study with eight subjects on disyllabic words. The results show that the question intonation in GuanM has a higher and raised register, compared to the statement intonation, given higher F0 mean and higher F0 maximum and minimum in questions. In addition, the longer duration of the last syllable of questions but not the first syllable correlates with the established definition of boundary tones, which typically occur on the last syllable of the intonational phrase. Finally, in the T2T2 question intonation, the F0 change rate in the second syllable was higher, while in T1T1, it was lower, compared to statement counterparts, highlighting the necessity and a characteristic of the high boundary tone (H%) in the question intonation of Chinese dialects that it moderates the extent of the falling tone and facilitates the rising tone rising.

Unraveling Students’ Liking of Teachers: The Impact of Multimodal Cues during L2 English Vocabulary Teaching

Jing Zhou and Yan Gu

While speakers’ wording, prosody and gestures may affect perceivers’ liking of speakers, few studies investigate how teachers’ multimodal cues jointly impact students’ evaluation of second language (L2) teaching. We extracted 54 videos of vocabulary instruction delivered by four female L2 English teachers, varying features of prosody, linguistics, and gestures. 156 university students randomly watched 12 videos and rated their liking of each vocabulary teaching. Prosodic (speaking rate, mean pitch), linguistic (utterance length, question rate, total words) and gestural cues (iconic, beat) of videos were coded and analysed as predictors, while controlling for different teachers, teachers’ dressing formality, students’ working memory, English proficiency, and familiarity with the target vocabulary. Results showed that better working memory, higher English proficiency, and prior knowledge of the target word were positive predictors of students’ liking of teaching. Teachers using longer utterances, asking more questions tended to be less liked by students. Furthermore, male students significantly preferred teaching with a slower speaking rate, lower mean pitch, higher iconic gesture rate but lower beat rate, and more formal teacher attire. However, these effects were not significant for female students. In conclusion, teachers’ multimodal cues influence students’ liking of L2 teaching, with implications for education practice.

Inter- and Intra-speaker Variation in the Acoustic Realizations of Quzhou Wu Tones

Lei Wang, Xinyue Yao, Bijun Ling and Xinlu Yang

The present investigation presents a multi-speaker acoustic study on the citation tones in Quzhou Wu. Ten native speakers participated in a production experiment. Fine-grained phonetic data documented the inter- and intra-speaker variation in the f0 contours of the tone categories. The results showed that inter-speaker variation is present in all tone categories and inter-speaker variation is greater than intra-speaker variation. The present investigation contributes to both the phonetic description of languages with complex tone systems and the documentation of inter-individual within-category variation.

A preliminary study on tonal variations in Singapore Teochew

Jiajia Cai

This study examines the phonetic properties of lexical tones in Singapore Teochew (SgT). An acoustic analysis of 18 speakers' tone production using generalised additive modelling reveals variations in both pitch height and contour shape for SgT compared to the variety spoken in China. Specifically, the mid-level T1 and high-level T2 are generally flatter and exhibit a more level contour in SgT. The high-rising T4 in the Singapore variety is significantly lower and displays a dipping contour not found in Standard Teochew. Additionally, the low-dipping T5 and low-level T6 are merged into a falling tone in SgT. By comparing the tone production of SgT speakers with that of English-Mandarin bilinguals, this study suggests that transfer from Singapore Mandarin is likely to be a significant factor in shaping the overall tonal variations in the local Teochew variety.

Incomplete neutralization in tone sandhi in Taiwan Southern Min spontaneous speech

Yu-Ying Chuang and Sheng-Fu Wang

In Taiwan Southern Min, a syllable's underlying tone is only realized in phrase-final positions, whereas in non-boundary positions, it is realized as another tonal category due to the application of sandhi rules. Previous studies found that the realizations of sandhi tone (e.g., high-falling becomes high-level) are indistinguishable from those of surface base tone (e.g., high-level realized as high-level), suggesting tonal neutralization. However, given that past investigation utilizes mainly laboratory speech, and that with spontaneous speech data, the control for confounding factors is insufficient, it remains unknown whether neutralization is genuine, and whether neturalization interacts with speech register. The goal of the present study is to examine sandhi tone realizations in spontaneous speech in greater detail. We analyzed the entire F0 contours of monosyllabic words with Generalized Additive Models. Specifically, for every tonal category, we compared the realizations of sandhi tone with their respective base tones. Results showed that sandhi and base tones differ in not only pitch height but also contour shape, though to a varying degree depending on gender and tonal category. The current finding provides evidnece of incomplete neutralization in spontaneous speech, hence posing an interesting question to the phonological status of tone sandhi in Taiwan Southern Min.

Baby Register and Adult Register in Infant-directed Speech

Tae-Jin Yoon and Seunghee Ha

This study aims to present a comprehensive description of the acoustic features of infant-directed speech (IDS) compared to adult-directed speech in natural home environment. This study involved 11 pairs of 9 to 12-month-old Korean infants and their parents, selected from a larger study. The infants had no medical or developmental issues, and normal hearing. Data collection involved home visits, parental interviews, and the use of LENA recorders to capture infants' language environments. Acoustic analysis focused on fundamental frequency, intensity, and harmonic-to-noise ratio. Statistical tests were conducted to compare adult-directed speech (ADS) and infant-directed speech (IDS) acoustic properties, assess feature distribution, and quantify the overlap between distributions. The analysis aimed to understand how speech properties differ in interaction with adults and infants. The study hypothesizes distinctive acoustic features in Korean IDS and proposes a dynamic construct with two registers. The authors employ statistical analyses and density plots to assess the distribution of acoustic features, challenging the conventional notion of IDS as a monolithic entity. The results of this research underscores the dynamic nature of Infant-Directed Speech (IDS) in the Korean context. Contrary to traditional views, the study introduces the concept of two registers within IDS – Baby Register and Adult Register.

Same Sentences Different Meanings: Prosodic and Gestural Resolution of Ambiguity in Mandarin Chinese

Jiajun Gao and Yan Gu

Speakers use prosody to resolve ambiguity, but what if prosody cannot make distinctions? We explore (1) how speakers employ prosodic and gestural cues to deal with sentences with ambiguous meanings and (2) what insights the audiovisual resolution of ambiguities offers regarding communicative efficiency and effort. Thirty-two native Chinese speakers were asked to articulate twenty-two ambiguous Mandarin sentences. Half could be semantically differentiated using prosody, and half could not. Firstly, participants articulated all ambiguous sentences spontaneously and provided explanations to a confederate, revealing their dominant interpretations. Secondly, participants articulated the same ambiguous sentences twice, each time guided by a hint suggesting a different meaning. Participants' prosodic cues and gestures were coded and analyzed. The results showed that for ambiguous sentences that can be prosodically distinguished, participants employed various prosodic cues such as pausing, tones, stress, and speaking rates. Additionally, 51.85% of sentences were accompanied by referential (iconic; pointing) gestures, while 17.33% of sentences were accompanied by non-referential (beat; interactional) gestures. However, when prosodic cues were unable to mark ambiguity, participants resorted to more referential gestures (97.30%) but fewer non-referential gestures (1.28%). In conclusion, speakers adopt a multimodal approach to enhance communicative efficiency while there is a trade-off between modalities.

Trading Relations in Segmental Cues to Prosodic Prominence

Shawn Foster and Jennifer Cole

English vowel formants undergo phonetic enhancement under prosodic prominence. However, vowel classes differ in the extent and dimensions of enhancement in ways that remain poorly understood. This study investigates differences in the effect of prosodic prominence on the production of eight English vowels. Thirty speakers each produced 48 repetitions of nonce words containing critical vowels. To elicit prominence, speakers produced short phrases that either repeated after a model talker or corrected the talker's use of one word.

Bayesian multivariate mixed-effects models were used to assess the effect of prominence and vowel on F1 and F2. Results suggest an inverse relationship between vowel height and effect of focus on F1, as well as a trade-off between the use of F1 and F2 to signal prominence. For the lowest vowels examined, /ɛ/, /ʌ/ and /a/, prominence was associated with raised F1 and no accompanying movement of F2. Conversely, the high front vowels /i/ and /e/ peripheralized along F2 under prominence without moving along the height dimension. Vowels intermediate to these showed movement along both dimensions. We discuss these results in terms of our understanding of the relationship between the expression of prominence and the phonological specification of vowels.

Role of Mispronunciation of Pitch Accent in Lexical Access in Japanese

Terumichi Ariga

While previous studies have shown that pitch accent constrains lexical access in Japanese, the role of mispronunciation of it is less often addressed. Two types of pitch accent mispronunciations exist in Japanese. One implies that mispronouncing results in other lexical items; when jidoo HLL "children" is prosodically mispronounced as jidoo LHH, it is another real word ("automaticity"). The other implies that mispronunciation may result in nonwords; when tansu LHH "wardrobe" is mispronounced as tansu HLL, it is a nonword. The present study investigated the process of lexical access to prosodically mispronounced words using a priming paradigm. When a prosodically mispronounced nonword (e.g., tansu HLL as a mispronunciation of tansu LHH "wardrobe") was presented as a prime, the response to a semantically related target word was delayed compared to when the prime was a semantically unrelated control word. However, when a segmentally mispronounced nonword (e.g., dansu LHH as a mispronunciation of tansu LHH) was presented, such a delay was not observed. These results indicated that the mispronunciation of pitch accent interfered with lexical access more significantly than the mispronunciation of a segment, supporting the view that pitch accent plays a crucial role in lexical access in Japanese spoken word recognition.

Focus Prosody in Shanghai Chinese

Jingwen Huang, Aijun Li and Zhiqiang Li

Shanghai Chinese exhibits distinct left-dominant (LD) and right-dominant (RD) tone sandhi patterns at the word and phrasal levels. This study examines the prosody of various focus structures in declarative and interrogative sentences of the Subject-Modifier-Verb-Object type, focusing on interaction between tone sandhi and intonation when target words receive focus. The results show that narrow focus (NF) leads to nuclear accent on the target words while the sentence stress predominantly manifests on the object in the context of broad focus (BF). Narrowly focused words positioned initially or in the middle of sentences show significantly extended pitch range and duration in comparison to those in BF. On the target words, the first syllables typically display a wider pitch range than second syllables in BF, but the second syllables often match or exceed the pitch range of the first in NF. In the sentence-final position, BF and NF produce no divergence in pitch and duration on target words, although tonal durations preceding target words in NF tend to be longer than those in BF. LD and RD patterns respond to focus prosody and intonation boundary differently, with the RD pattern observed more often in NF.

Using role-playing tasks to document intonational tune prototypes in Nasal, an endangered language of Sumatra

Jacob Hakim

This paper describes a scripted role-playing task used to elicit a basic inventory of intonational tune types in Nasal (ISO 639- 3: nsy; glottocode nasa1239), an endangered Austronesian language of Sumatra currently spoken in three villages by around 3,000 people. This study is a subset of the ongoing prosodic description that forms part of a larger Nasal documentation project. Prosodic description is not often included in language documentation, especially for Austronesian languages; when included, these descriptions are often very limited or based on impressionistic descriptions (Himmelmann & Kaufman 2020). I make the case here that prosodic description based on carefully planned experimentation and acoustic measurement is not only achievable but a necessary part of linguistic fieldwork for language documentation. In the case of Nasal, eight participants (four men and four women) were recorded reading scripted lines from role-play dialogues in a variety of real-life scenarios. These recordings were transcribed, labeled according to the HCRC dialogue coding scheme (Anderson et al. 1991, Carletta et al. 1996), and analyzed with a cluster analysis using the Contour Clustering app (Kaland 2023). The alignment of tune patterns with different utterance types reveals a preliminary inventory of prototypical intonational tunes, including a falling tone pattern for polar questions, a pattern that is crosslinguistically uncommon.

Use of Word-Level Stress in L2 Spanish Word Recognition

María Teresa Martínez García, Julie Kamber and Sandra Schwab

Individuals with different L1s rely differently on suprasegmental cues when recognizing spoken words in L1 and in L2. The present study investigates the use of stress information in L2 word recognition and includes intermediate-level learners of Spanish with different L1s: German (i.e., a language with distinctive word stress; DE), French and Korean (i.e., languages without distinctive word stress; FR, KR). Contrary to DE, the absence of word stress in French and Korean was expected to hinder FR and KR ability to use stress in L2 Spanish word recognition. In a cross-modal word-identification task, participants listened to semantically ambiguous auditory sentences ending with incomplete two-syllable word fragments and had to choose the word that matched the heard fragment, which either contained (stressed condition) or lacked a stressed syllable (unstressed condition). While the results revealed no difference between three language groups in the unstressed condition (62%), the groups differed in the stressed condition. KR accuracy (77%) was unexpectedly higher than FR (63%) and as good as DE (76%), suggesting that, contrary to FR, KR were able to use stress to access L2 words, although word stress does not exist in their L1.

Production and perception of emotional intonation among preschool children with cochlear implants

Fang Zhang, Hongtao Li and Ao Chen

Children with cochlear implants (CI) often have difficulties processing low frequency due to insufficient resolution of frequency filters of the CI, leading to inaccurate pitch perception. The current study investigated intonation perception and production of Chinese preschool children (3-5-year-olds) with CI. In Experiment 1, 55 children with CI and 44 children with normal hearing (NH) were auditorily presented with six pairs of semantically neutral sentences, with each pair including a happy and a sad intonation. The children were asked to identify the emotion of each sentence by selecting cards with corresponding facial expressions. Neither the NH children nor the children with CI were able to identify the emotions accurately. In Experiment 2, using a sentence repetition paradigm, another 51 children with CI and 32 NH children produced the six pairs of sentences used Experiment 1. Compared to the happy sentences, both groups showed a lower mean f0 for the sad sentences. Nevertheless, only the NH children made use of f0 range and intensity cues when differentiating the two emotions. These results indicate late development of perceptual emotional intonation representation, and the CI children made use of less cues compared to NH children when marking sentential emotion.

A Comparison of Synthesis Method Impact on Listener Perception of Play-Acted Speech

Emily Lau, Brechtje Post and Kate Knill

There has been an increased interest in both Linguistics and Artificial Intelligence research in play-acted expressive speech and its acoustic and perceptual characteristics, which are nuanced and difficult to define. This work compares the results of two sets of listening experiments that test the impact of the Bio-informational Dimensions (BIDs) on perceptions of play-acted speech using stimuli that were re-synthesised using different methods. One method performs pitch manipulations on each pitch point simultaneously, while the second separates these manipulations into separate steps. In both tests, participants listened to pairs of utterances that were resynthesized along the BIDs of size projection and dynamicity to varying degrees to simulate dramatic expressions of anger, and then rated the utterances' differences in dramatic expression. Size projection was found to have significant positive impact in both experiments. However, dynamicity had a slightly significant negative effect in the first experiment but no significant effect in the second experiment. These results prompt further questioning about the specific parameters that impact perceptions of vocal expression and those that should be targeted when synthesizing specific speech styles.

Cortical tracking of prosody after stroke and in aging: preliminary evidence from magnetoencephalography

Giada Antonicelli, Nicola Molinaro, Patricia De La Riva, Raquel Laspiur, Arantza Lopez de Turiso, Maddi Carrera and Simona Mancini

Evidence exists that brain-damaged but also healthy aging people exhibit difficulties in linguistic (LP) and emotional prosody (EP) and poorer neural synchronization to speech (cortical tracking of speech, CTS) in the delta/theta frequency band. Using a cross-sectional design with left-hemisphere (LH) and right-hemisphere (RH) stroke survivors, young (18-30 y.o) and old (35-80 y.o.) control participants (YC, OC, respectively), we ask whether: 1) CTS is anomalous after stroke and in healthy aging, 2) correlates with prosody interpretation, and 3) LP and EP processing are segregated in the brain. Participants listen to Spanish sentences in EP, LP, and neutral prosody conditions. Data from 8 YC, 4 people with LH damage (LHD), and 3 OC show that, relative to controls, LHD have lower task accuracy, and their CTS is lower in temporal areas in the delta/theta bands but higher in the delta band over parietal sensors. So far, we confirmed that delta/theta CTS is anomalous and prosody comprehension is harder after stroke, while no effects of prosody type were observed. It might be that since LHD experience a deficit in early input-driven processes, they need to recruit more top-down cognitive resources, which does not lead to control-like task performance.

Comparing the imitation of naturally-produced and synthesized F0 in American English nuclear tunes

Jeremy Steffman and Jennifer Cole

Imitation tasks are used in intonation research to identify properties that are perceptually salient and encoded for subsequent production. The current study examines whether and how imitation of synthetic versus naturally produced F0 contours may differ. We compared F0 contours in American English from two imitation experiments where participants heard sentences with the same phrase-final intonation and reproduced the heard pattern on a novel sentence. In one experiment, F0 patterns of stimuli were controlled via pitch resynthesis using straight-line approximations of (phonological) tonal targets; the other used natural productions of the same tunes. F0 trajectories were examined to identify which F0 properties of the stimuli were preserved or lost as a function of the type of stimulus. Imitations of natural vs. resynthesized stimuli were compared using time-series k-means clustering analysis, GAMM modeling, and RMSD as a measure of F0 difference between imitation and stimulus. We observe striking similarity in imitations of natural and resynthesized stimuli based on clustering solutions, with small, localized differences in GAMMs for only two out of eight tunes tested. RMSD results show closer imitation with resynthesized stimuli, suggesting greater attention to fine phonetic detail of F0 patterning when other cues to intonational contrasts are held constant.

Music in the treatment of childhood speech sound disorders: Evaluating prosody in Dutch-speaking children

Mirjam van Tellingen, Joost Hurkmans, Hayo Terband, Ben Maassen and Roel Jonkers

Purpose: Music is frequently used in the treatment of childhood speech sound disorders (SSD). The prosodic overlap between speech and music is explored to improve both segmental accuracy and prosody in children with SSD. In a pilot study, evaluation of Speech-Music Therapy for Aphasia showed improved speech production at the level of segmental accuracy and intelligibility. However, measures of fluency used in that study were inadequate for evaluating prosody.

Method: Two new methods for the evaluation of prosody were developed. The first method was a scale for perceptual judgement of prosody in spontaneous speech through a focus on naturalness. This scale was evaluated and validated in a group of children with and without SSD. The second method was a task for the realisation of lexical stress in non-words.

Results: Preliminary results of the evaluation of the prosody judgement scale indicate that the scale discriminates between children with and without SSD.

Conclusion: The evaluation of prosody in treatment studies of children with SSD was not possible with existing measures. A newly developed prosody judgement scale is suitable for the assessment of prosody in spontaneous speech. Evaluation of the developed lexical stress task is ongoing. Preliminary results will be ready for presentation.

The assessment of automated rating of L2 Mandarin prosody in lexical tone recognition and pauses

Yao Wu

Language instructors spend a lot of effort providing feedback on spoken language tasks. The use of automated engines can be helpful in improving pronunciation for tonal languages like Mandarin. The study focused on evaluating prominent APIs for their ability to assess oral readings by L2 Mandarin learners, with particular emphasis on detecting lexical tones and pauses.

In evaluating the recognition of lexical tones, the automated engines demonstrated the capacity to detect tones and provide a corresponding tonal pronunciation score. The study found that the overall accuracy of tonal diagnosis reached 80%, as determined through the calculation of false rejection and false acceptance rates. Furthermore, the rating distribution of the four lexical tones closely aligned with human ratings, thereby offering valuable insights for instructional strategies aimed at learners.

Regarding the assessment of intonation and other prosodic features, the automated engines were able to generate scores for pauses and overall fluency. It was observed that these scores exhibited a correlation of 0.6 with human raters’ detection of unnatural prosodic boundaries. As a result, the study recommends the inclusion of more detailed descriptions of prosodic features in current Mandarin automated rating models to enrich feedback and learning opportunities for learners.

Chinese EFL Learners' Perception of English Emotional Prosody

Yanyang Chen and Ying Chen

Chinese EFL learners' perception of English emotional prosody was investigated for the effects of emotion type (neutrality, happiness, surprise, sadness, and anger), speech condition (normal speech and low-pass filtered speech), learner's English proficiency (low-proficiency and high-proficiency), and learner's gender (male and female) via forced-choice identification tasks. Eighty native speakers of Mandarin, who learned English as a foreign language (EFL) in China, were requested to identify normal and filtered speech of five basic emotions produced by ten native speakers of American English. Results indicated that all the five emotions were recognized at rates above chance with negative expressions (i.e., sadness and anger) more accurately and promptly identified than positive expressions (i.e., happiness and surprise). Normal speech yielded more accurate identification than filtered speech, except for the expressions of surprise. Furthermore, the identification accuracy of learners with high English proficiency for happiness was significantly higher than that of low-proficient learners. Female learners' identification accuracy was marginally higher in negative emotions than males. This study pedagogically suggests explicit instructions on English emotional prosody, particularly positive emotions, to Chinese EFL learners with special attention to male and low-proficient learners, who seemed to be less sensitive to English emotional prosody than female and high-proficient learners.

Incremental Processing of Prosody in L2: A Visual World experiment with French learners of English

Chie Nakamura, Hiyon Yoo and Giuseppina Turco

Native listeners rely on prosodic cues for the resolution of syntactic ambiguity at very early stage of online sentence processing. In the current study we test whether a similar mechanism is shared by second language (L2) listeners. In a visual word paradigm experiment, we used sentences with PP attachment ambiguity such as The boy will write to the panda with the crayon and tested French learners of L2 English. We examined the impact of the prosodic boundary that was placed either before or after the patient NP (e.g., the boy will write to % the panda with the crayon, or the boy will write to the panda % with the crayon). Results show that learners are able to integrate prosodic boundary information for the resolution of syntactic ambiguity but at a delayed time than natives. These findings corroborate previous work testing a different L2 population (i.e. L1-Japanese L2-English), thereby suggesting that general mechanisms drive parsing decisions for L2 learners, even when L1/L2 pairs exploit similar prosodic cues to locate boundary information.

Bai tone perception and production by Naxi speakers in Jiuhe: a preliminary study

Meihao Wan and Peggy Mok

This study investigated the perception and production of Bai tones by native Naxi speakers in Jiuhe. Jiuhe Bai features six lexical tones distinguished by pitch and phonation, while Jiuhe Naxi has three tones differentiated only by pitch. We explored whether Naxi speakers could accurately perceive and produce the Bai tones, particularly those with similar pitch contours. Ten native Naxi speakers participated in the perception and production experiments. The perception experiment involved a discrimination task with all possible pairs of Bai tones. In the production experiment, participants were instructed to produce all the Bai tones with a wordlist of minimal pairs. Both acoustic and EGG signals were recorded and analyzed for pitch and phonation patterns. The discrimination results revealed that native Naxi speakers had difficulty distinguishing certain Bai tones, which were also merged in their own production. Interestingly, not only the tones with similar pitch contours but also those with similar phonation patterns are prone to confusion among the Naxi speakers. The findings can shed new light on the acquisition of L2 tone categories in nonnative speakers, specifically in languages that employ multiple cues for tone distinction.

Language redundancy effects on the prosodic word boundary strength in Standard German

Tianyi Zhao, Tina Bögel, Alice Turk and Ricardo Napoleão de Souza

The Smooth Signal Redundancy hypothesis proposes a complementary relation between language redundancy and acoustic redundancy mediated via prosodic prominence and boundary structure. To test this hypothesis, the current study investigates the effects of lexical frequency on boundary-related segmental (+pause) duration patterns at prosodic word boundaries in Standard German. Results are consistent with predictions made by the Smooth Signal Redundancy hypothesis, showing an inverse correlation between lexical frequency and duration: Word boundary-related target intervals for frequent words were more than 10% shorter than corresponding intervals for infrequent words, and effects on non-boundary related intervals were not significantly different for frequent vs. infrequent words. These effects suggest a preference for producing stronger prosodic boundaries in case of low language redundancy. The effects appear to be stable across varying speech tempo by different speakers and targets, even when the factor of speech tempo is controlled for. This is consistent with the view that speech tempo, as a global factor that modulates the overall utterance, does not interfere with the localized acoustic redundancy.

Semantic priming and prosodic structure: At the interface between language redundancy and acoustic salience

Mila Freiseis, Tianyi Zhao and Tina Bögel

The Smooth Signal Redundancy Hypothesis (SSRH) states that there is an inverse relationship between language redundancy and acoustic saliency. Less redundant items, e.g. infrequent or unpredictable ones, become more salient, and vice versa. The SSRH further assumes that prosodic structure, i.e. prosodic boundaries and prominence, mediates the relationship between language redundancy and acoustic saliency. In this paper, we tested whether semantic priming, one of the measures of language redundancy, affects prosodic structure at word boundaries. In a production experiment we presented German sentence pairs with identical target words. These target words were presented either in a context where they were primed by semantically related words or in a context where they were not primed. Results showed an effect of semantic priming on prosodic structure in that primed targets were significantly shorter than non-primed ones. This effect was increased when measures of lexical frequency were taken into account as well.

Sound effect, onomatopoeia, and iconic prosody in Chinese: Emerging vocal iconicity in child-directed speech and child production

Mengru Han, Yiqi Nie and Yan Gu

Iconicity plays an important role in language acquisition and cognition. This study aimed to better understand the use of three types of vocal iconicity in language input and child production: sound effects (e.g., making the sound of eating), onomatopoeia (e.g., meow), and iconic prosody (e.g., faaar). We coded these aspects in a corpus of Chinese adult-directed speech (ADS) and child-directed speech (CDS), in which mothers semi-spontaneously told the same story to an adult and their 18-month-old (N = 21) or 24-month-old (N = 19) children. We examined whether mothers' vocal iconicity differs between CDS and ADS and how it emerges in child production. We found that (1) mothers used significantly more sound effects and iconic prosody, but not onomatopoeias, in CDS compared to ADS; (2) In CDS, the proportions of the three types of iconicity ranked as iconic prosody>sound effects>onomatopoeias, whereas the proportions for children emerged as sound effects>iconic prosody and onomatopoeias; (3) Chinese children aged 18 or 24 months produced little onomatopoeia and iconic prosody (except for one instance at 24 months). In conclusion, iconicity is more prevalent in CDS than in ADS, and iconic prosody is an advanced prosodic skill that is not typically developed by two-year-old children.

CF0 effect and articulatory strength of geminate consonants

Sireemas Maspong, Francesco Burroni and James Kirby

This study explores how fundamental frequency (F0) and articulatory strength are related in Italian geminate consonants. Consonant-intrinsic F0 (CF0) effects are examined with a focus on the hypothesis that geminates exhibit such effects as a consequence of their inherent “tense” articulation, manifested as higher F0 and a more constricted articulatory target. Simultaneous articulatory and acoustic data were collected from 10 native Central and Southern Italian speakers pronouncing six disyllabic nonce words ([ip(ː)a, ib(ː)a, im(ː)a]) within a carrier sentence varying in speech rates. F0 values were extracted at 10 ms intervals before and after consonantal closure, while articulatory data, including Minimum Lip Aperture (LA) and Maximum Jaw Height (JH), were recorded using an AG501 Carsten EMA. Linear mixed-effect regressions were fit to the data. The findings reveal that geminates exhibit higher post-closureF0, lower LA, and higher JH compared to their singleton counterparts, supporting the hypothesis that geminate consonants possess “tense” properties. Additionally, weak positive correlations were observed between post-closure F0 and LA.

Gradiency and categoriality in the prosodic modulation of French Sign Language: A kinematic approach using Electromagnetic Articulography

Justine Mertz, Lena Pagel, Giuseppina Turco and Doris Mücke

During interaction speakers tend to adjust the amount of coarticulatory cues to increase or decrease perceptual distances between competing speech units. Anticipatory coarticulation has also been observed in the visual-gestural modality. Despite this, little is known about the use of coarticulatory strategies in sign language. The present study is the first to investigate coarticulation in French Sign Language (LSF) using 3D Electromagnetic Articulography (EMA) to provide precise kinematic measurements in sign production. In this novel approach, a deaf native signer was recorded (EMA/video) producing phonological pairs of signs composed of '1'- and/or '3'-handshape. Our findings demonstrate that kinematic data allows for the detection of coarticulation in various discourse contexts. Temporally, we observe the anticipation of the '3'-handshape before the end of its immediately preceding '1'-handshape sign (and vice versa). Spatially, the (repetitive movement of) the sign is affected by reduction/truncation if followed by another sign. Within a dynamical approach (Articulatory Phonology), we analyze the kinematics of our sign data as a result of systematic patterns of overlapping organization triggered by the phonological system. Based on this view, we attempt to take a step forward towards an integration of gradient and categorical processes such as coarticulation and assimilation.

When “uhm”, “and” and “yeah” sound the same — prosodic aspects of discourse pragmatic markers in American English

Marlene Böttcher and Margaret Zellers

This study explores the distributional and prosodic similarities and differences of different particles in discourse. Both lexicalized discourse particles (e.g., yeah, so, well) and filler particles (e.g., uh and uhm) share structuring discourse functions in English. This study looks at different types of particles and investigates their occurrence at the boundary between broader units of discourse, where they are likely to be used in a structuring function to indicate so called frame shifts.

The data analysis focuses on elements occurring at discourse boundaries in 40 formal and informal English narrations by mono- and bilingual speakers from the RUEG corpus. Since prosody is also an important means in discourse organization, the analysis also includes prosodic aspects of the discourse particles.

A variety of elements and their combinations was found at discourse boundaries, including filler particles, discourse markers, tongue clicks and connectors. While filler particles, alone or in combination, are more frequent in formal narrations, discourse markers and connectors are more frequently found in informal narrations. Prosodically, the boundary elements were produced similarly in pitch and duration, but different types also showed finer phonetic differences. Both the choice of particle and their prosodic realization are influenced by the formality of the situation.

Exploring the role of personality traits in the imitation abilities of non-native speech in familiar and unfamiliar languages

Peng Li, Ioanna Ioannidou, Ilaria Marazzina, Paula Pericacho, Béibhinn Reardon and Lu Xing

Previous studies have explored the influence of sociopsychological factors on second language (L2) pronunciation, yet personality traits remain relatively underexplored in this context. Notably, the interplay between speakers’ familiarity with the target L2 and the predictive role of personality traits in L2 speech production has not been thoroughly investigated. This study used a speech imitation task to assess the speech production abilities of 35 L2 speakers of English, who had no prior knowledge in Chinese, in both English (familiar L2) and Chinese (unfamiliar L2). Native speakers of English and Chinese rated the accuracy of sentences imitated by the participants. Personality traits were evaluated using the Multicultural Personality Questionnaire (short form) in five aspects: cultural empathy, flexibility, social initiative, open-mindedness, and emotional stability. The findings from a linear mixed-effects model revealed that only cultural empathy showed a significantly negative effect on speech imitation scores for Chinese, not for English. This implies that good cultural empathy may hinder individuals from accurately imitating the accent of an unfamiliar L2, possibly due to concerns about cultural misunderstandings. These results suggest that certain sociopsychological measures may pose challenges to initial encounters with a novel language

The perception of Spanish lexical stress by proficient Mandarin learners of Spanish

Peng Li and Xiaotong Xi

Unlike Spanish natives, Mandarin speakers tend to produce the Spanish lexical stress contrast by manipulating pitch rather than other prosodic cues such as duration. However, the perception of Spanish lexical stress remains less clear. This study examines Mandarin speakers' cue-weighting strategies in perceiving Spanish stress and investigates how musical perception aptitude and auditory processing abilities affect cue-weighting. Twenty-two L1 Mandarin speakers with advanced Spanish proficiency and 19 Spanish natives participated in a stress perception task, which involved identifying strong-weak and weak-strong lexical stress patterns. The pitch and duration ratio of the target word's vowels were manipulated (7 steps each). Musical perception aptitude (i.e., accent, melody, rhythm, and pitch) and auditory processing abilities (i.e., duration and pitch) of Mandarin speakers were also assessed. Results show that Mandarin speakers rely more on pitch than duration to identify Spanish lexical stress patterns, in contrast to Spanish natives (duration > pitch). Additionally, only musical accent perception skills significantly predicted Mandarin speakers' cue-weighting, with higher musical accent scores correlating to larger weighting on pitch cues. The results suggest that even advanced learners exhibit L1 transfer in prosody to L2, and musical perception skills may play a role in L2 prosodic acquisition.

Speech rate and prosodic phrasing interact in Korean listeners' perception of temporal cues

Jeremy Steffman, Sahyang Kim, Taehong Cho and Sun-Ah Jun

This study explores the interaction between contextual speech rate and prosodic phrasing in listeners' perception of temporal cues in Korean. We investigate perception of the aspirated/fortis stop contrast, testing categorization of a Voice Onset Time (VOT) continuum. Aspirated stops have longer VOT and shorter vowel duration relative to fortis stops. Vowel duration in both stop categories is lengthened at the beginning of a prosodic phrase, and this prosodic strengthening pattern has been shown to influence perception of the stop contrast with temporal context controlled. Building on this previous finding, we manipulate preceding speech rate in a carrier phrase (slower/faster) and cross this with a phrasing manipulation: 1) no prosodic juncture before the target, 2) a preceding intonational phrase boundary (cued by pre-boundary lengthening), and 3) the same boundary with an additional pause. Results show canonical speech rate effects only in the absence of a preceding boundary. Prosodic strengthening effects, which show additive differences based on boundary strength, are present only when speech rate is slow. In sum, findings suggest that speech rate effects are influenced by prosodic phrasing, and phrasing effects are influenced by speech rate, providing insight into the interplay of these factors in shaping temporal cue perception.

Investigating the role of semantics and perceptual salience in the memory benefit of prosodic prominence

Yuxi Zhou, Constantijn L. van der Burght and Antje S. Meyer

Prosodic prominence can enhance memory for the prominent words. This mnemonic benefit has been linked to listeners' allocation of attention and deeper processing, which leads to more robust semantic representations. We investigated whether, in addition to the well-established effect at the semantic level, there was a memory benefit for prominent words at the phonological level. To do so, participants (48 native speakers of Dutch), first performed an accent judgement task, where they had to discriminate accented from unaccented words, and accented from unaccented pseudowords. All stimuli were presented in lists. They then performed an old/new recognition task for the stimuli. Accuracy in the accent judgement task was equally high for words and pseudowords. In the recognition task, performance was, as expected, better for words than pseudowords. More importantly, there was an interaction of accent with word type, with a significant advantage for accented compared to unaccented words, but not for pseudowords. The results confirm the memory benefit for accented compared to unaccented words seen in earlier studies, and they are consistent with the view that prominence primarily affects the semantic encoding of words. There was no evidence for an additional memory benefit arising at the phonological level.

A corpus phonetics study of Dalabon nouns

Catalina Torres and Sarah Babinski

Dalabon is a severely endangered Australian language in the Northern Territory. The language's intonational phonology has been described as a head-edge marking type. As several other Australian languages, Dalabon has been described as a stress language, based on a stress rule. However, there are no dedicated studies examining the acoustic cues involved in marking stress in this language. In this preliminary study, we investigate a set of acoustic correlates commonly associated with stress marking in other languages to examine their potential role in Dalabon. For this purpose, we use corpus data of spontaneous speech taken from personal narratives and elicited story telling. The data from the DoReCo corpus is transcribed, translated into English, time-aligned at the level of discourse units, and forced aligned at the segmental level. We further process the corpus to obtain a word list of nouns, a syllable level and predicted stress marking. In our acoustic analysis, we examine duration, fundamental frequency, the first and second formants, as well a relative intensity measured in vowels. Our statistical investigation shows that these cues don't associate with stress marking. Instead, we find that major prosodic boundaries have an effect on some of the cues.

Individual Variation in Phonetic Accommodation of Mandarin-Speaking Children during Conversations with a Virtual Robot

Yitian Hong and Si Chen

Recent studies on child-robot interaction (CRI) emphasize that children's speech behaviors towards robots are shaped not only by their beliefs about the robot but also by individual variations in how they perceive and build rapport with robots. Speech accommodation, characterized by adjusting speech features in response to the other talker, is a valuable indicator of child speech in CRI. While previous research mainly focused on simple interacting tasks, little is known about natural conversations between children and robots.

In our study, fifty-five Mandarin-speaking children collaborated with the virtual robot Furhat to identify differences between pictures using spoken language. Keywords were recorded before and after the interaction. Acoustic analysis revealed significant reduction of differences in fundamental frequency, vowel duration, and vowel formants of the keywords between the child and the robot. Importantly, their accommodation demonstrated substantial individual variabilities, guided by the child's personality and slightly influenced by their perception of the robot's 'agreeableness,' interpreted as its degree of human-likeness. This is the first study investigating speech accommodation in natural conversations between Mandarin-speaking children and a social robot. It provides new evidence supporting a hybrid model combining automaticity and social motivations for interpreting accommodation in child communication, Mandarin speakers, and human-robot interaction.

Deaccented Verb as an Element in the Utterance Information Structure

Jan Volín and Adléta Hanžlová

The chief objective of the present study is to investigate actual manifestation of the potential lexical stress in Czech verbs. Putatively, lexical stress is expected to materialize in all auto-semantic words of an utterance. However, due to contextual givenness and stress-clash rule effects, some of the words can be deaccented. To map the situation, continuous spoken texts rather than isolated sentences need to be examined.

Narratives produced by 16 professional speakers were annotated in terms of manifest accent-groups. In the recordings, 3708 verbs were identified and sorted into 5 grammatical classes. These were first inspected in a binary fashion: the structural stress either materializes or not (i.e., the verb is deaccented). Further descriptors of the verb status in the accent-groups configurations were extracted in order to find out how often the produced forms can be explained with reference to the context and how often various other factors were in force. Complementary questions concerned accent placement on auxiliary and modal verbs. The results offer an insight into a rich pool of pragmatic relations of verbs with other constituents, and provide a quantitative base for further experiments in the field of the information structure of utterances.

A phonological model of Atara Imere intonation

Adam Chong and Coppe van Urk

In this study, we provide a preliminary Autosegmental-Metrical model of the intonational phonology of Atara Imere (Polynesian; Vanuatu). Our initial analysis suggests that Atara Imere has three tonally marked prosodic units above the prosodic word: (i) Accentual Phrase (AP), (ii) Intermediate Phrase (ip), and (iii) Intonational Phrase (IP). We confirm previous observations that lexical prominence occurs on the antepenultimate mora, and this position is marked primarily with a LH* pitch accent. The right edge of AP boundaries are primarily marked by a L tone (La), though we also find evidence for a less frequent H boundary tone (Ha). So far, we have found evidence for one ip boundary tone: a high tone (H-) marking the right edge of larger syntactic units. Finally, IP boundaries are marked by either L% or H%. Focus and intonational patterns of different sentence types are also discussed.

Usefulness of Emotional Prosody in Neural Machine Translation

Charles Brazier and Jean-Luc Rouas

Neural Machine Translation (NMT) is the task of translating a text from one language to another with the use of a trained neural network. Several existing works aim at incorporating external information into NMT models to improve or control predicted translations (e.g. sentiment, politeness, gender). In this work, we propose to improve translation quality by adding another external source of information: the automatically recognized emotion in the voice. This work is motivated by the assumption that each emotion is associated with a specific lexicon that can overlap between emotions. Our proposed method follows a two-stage procedure. At first, we select a state-of-the-art Speech Emotion Recognition (SER) model to predict dimensional emotion values from all input audio in the dataset. Then, we use these predicted emotions as source tokens added at the beginning of input texts to train our NMT model. We show that integrating emotion information, especially arousal, into NMT systems leads to better translations.

Acoustic analysis of several laughter types in conversational dialogues

Kexin Wang, Carlos Ishi and Ryoko Hayashi

Previous studies suggest the existence of two distinct forms of laughter: mirthful/spontaneous laughter and social/intentional laughter. The current work aims to expand our understanding of the motives behind laughter and its functions in social conversation. About 1000 laughter events from 4 males and 4 females were extracted from multi-speaker conversation data, and the four predominant categories were used for acoustic analysis: mirthful, boosting, smoothing, and softening. Mirthful laughter and boosting laughter exhibit longer duration, higher F0 mean, intensity and HNR, as well as lower H1-A1 than other types, which suggest that laughter produced with positive emotion or attitude tends to have longer, higher and tenser voice quality. On the other hand, smoothing laughter and softening laughter displayed opposite characteristics, which indicates that intentional laughter emitted to smooth the interaction or soften the atmosphere can be acoustically identified to some extent from those with positive emotions. This work provides evidence that laughter with different functions has different acoustic characteristics that help us understand what laughter means in dialogue.

Contrast and predictability in the variability of tonal realizations in Taiwan Southern Min

Sheng-Fu Wang

Variability of speech signals is known to reflect linguistic units' predictability, which is often measured with lexical frequency and contextual probability. Since the same aspects of the signals often serve other functions such as conveying lexical contrasts and marking phrasal boundaries, it is interesting to examine to what extent the link between predictability and phonetic variability is constrained and manifested differently across different dimensions of phonological contrasts. This question motivates the present study's focus on F0 realizations in Taiwan Southern Min, a language with a rich inventory of lexical tones. From a corpus of spontaneous speech, the relationship between predictability measurements and the realization of different tonal categories is analyzed with Generalized Additive Models. Results show that predictability effects generally do not neutralize critical F0 contrasts (e.g., F0 peak between high-falling and low-falling tones, F0 mean between high-level and low-level tones). For some tonal pairs, there is a trend of lower predictability (i.e., higher surprisal and informativity) correlating with a larger F0 difference. These findings shed light on how the relationship between phonetic variability and predictability is modulated by the maintenance and enhancement of lexical contrasts in speech production.

Prosody can provide subtle disambiguating cues for local ambiguity resolution

Kathleen Schneider, Outi Tuomainen, Isabell Wartenburger and Sandra Hanne

The present study investigated the effects of prosodic cues for local ambiguity resolution in German SVO and OVS sentences. Thirty-two healthy participants were tested in a web-based two-alternative forced choice task to examine whether listeners are sensitive to two different prosody conditions for distinguishing SVO and OVS structures as quickly and accurately as possible. We examined a syntactically marked prosody condition (i.e., naturally produced f0 cues differentiating between SVO and OVS structures) and an enhanced prosody condition (i.e., marked prosody with naturally increased f0 maximum). Response accuracy and reaction times were assessed following signal detection theory and by running linear mixed models. We found only moderate discriminability of both word order structures with higher sensitivity levels for enhanced compared to marked prosody. This is in line with the mixed results of previous studies suggesting that prosodic cues constitute more subtle information for structural disambiguation of German SVO and OVS sentences. However, we add to those results by demonstrating a more facilitative role of enhanced prosody. Research on variability of prosodic word order cues in sentence comprehension still remains open to further investigation.

An acoustic-prosodic analysis of laughter types

Bogdan Ludusan, Marin Schröer and Petra Wagner

Laughter is a non-verbal phenomenon, widely used in human interaction, which has been shown to differ from speech along various acoustic-prosodic dimensions. Previous work has also revealed that the production of laughter is subject to a high degree of variation, with speakers normally having several types of laughter in their repertoire. Despite this, relatively little is known about how different types of laughter are marked prosodically and whether prosodic features may be used to discriminate between laughter types. We investigated here two types of laughter events produced in spontaneous interaction, with respect to five prosodic characteristics: duration, pitch, intensity, rhythm and voice quality. Our results showed that each of these characteristics, except for rhythm, differ between laughter types. We then employed prosodic features in a machine learning system trained to discriminate between three classes: speech and the two types of laughter. The proposed system obtained a similar speech/laughter classification performance to that of a system that considers only two classes, speech and laughter, while also having the advantage that a finer distinction, i.e., between laughter types, may be achieved.

The timing of beat gestures affects lexical stress perception in Spanish

Patrick Louis Rohrer, Ronny Bujok, Lieke van Maastricht and Hans Rutger Bosker

It has been shown that when speakers produce hand gestures, addressees are attentive towards these gestures, using them to facilitate speech processing. Even relatively simple "beat" gestures are taken into account to help process aspects of speech such as prosodic prominence. In fact, recent evidence suggests that the timing of a beat gesture can influence spoken word recognition. Termed the manual McGurk Effect, Dutch participants, when presented with lexical stress minimal pair continua in Dutch, were biased to hear lexical stress on the syllable that coincided with a beat gesture.

However, little is known about how this manual McGurk effect would surface in languages other than Dutch, with different acoustic cues to prominence, and variable gestures. Therefore, this study tests the effect in Spanish where lexical stress is arguably even more important, being a contrastive cue in the regular verb conjugation system. Results from 24 participants corroborate the effect in Spanish, namely that when given the same auditory stimulus, participants were biased to perceive lexical stress on the syllable that visually co-occurred with a beat gesture. These findings extend the manual McGurk effect to a different language, emphasizing the impact of gestures' timing on prosody perception and spoken word recognition.

Speech markers of Cancer-Related Cognitive Impairment: A pilot study

Amélie B. Richard, Alexandre Foncelle, Fabrice Hirsch, Sophie Jacquin-Courtois, Karen T. Reilly and Manon Lelandais

Speech is sensitive to mild cognitive changes due to age-related diseases, and prosodic features can identify patients with early-stage dementia from controls. Few studies have investigated speech markers of subtle cognitive impairment in non-neurodegenerative pathologies in younger populations, such as Cancer-Related Cognitive Impairment (CRCI). Little is known about the cognitive mechanisms underlying CRCI, but it is frequently encountered by cancer patients who mainly report memory-related concerns (i.e., forgetting words). Despite its substantial impact on patient quality of life, CRCI is difficult to detect with neuropsychological tools and often remains underdiagnosed. Our aim is to test whether previously documented speech markers are likely to detect CRCI in patients with breast cancer. We compared speech rate, F0 variability and pause duration in 11 breast cancer survivors with a cognitive complaint, 11 breast cancer survivors without any cognitive complaint and 10 controls in two narrative tasks (memory-based; picture-based). A Bayesian analysis showed no significant effects of group or task, but a qualitative analysis of pauses allowed us to generate hypotheses about the cognitive mechanisms underlying the patients’ reported memory concerns. Even though speech markers specific to CRCI have yet to be defined, prosodic analysis is a promising approach for detecting subtle cognitive impairment.

Effects of task type and task difficulty on oral fluency in native and non-native speech

Jörg Peters, Marina Frank and Tio Rohloff

The aim of this study was to explore differences in oral fluency between native and non-native speech, with a focus on the influence of task type and task difficulty. To reduce the impact of language structure on variability, the study compared High German (HG) and Low German (LG), two closely related languages with similar phonology, grammar, and vocabulary. Native speakers of HG, who had successfully completed a language course in LG, performed eight speaking tasks in both languages. To evaluate the effect of task difficulty on fluency parameters, three of these tasks were presented at different levels of task complexity, which was achieved by varying the availability of relevant information, the pre-task planning time, and the familiarity of the task. Measures of speed and breakdown fluency were obtained from both languages. As expected, LG speech showed lower speed and breakdown fluency compared to HG speech, but this effect varied by task type and task difficulty. We conclude that the assessment of oral fluency through effective variation of task type and task difficulty remains a major challenge for future research.

Intonational Patterns under Time Pressure: Phonetic Strategies in Bulgarian Learners of German and English

Judith Manzoni-Luxenburger, Bistra Andreeva and Katharina Zahner-Ritter

Research on the second-language (L2) acquisition of intonation is a growing field but only few studies have (so far) focused on the fine phonetic detail of intonational patterns in the L2. The present study concentrates on the phonetic realization of nuclear intonation contours under time pressure, testing Bulgarian learners in their L2s German and English – two languages in which intonation contours are accommodated differently by native speakers (L1) when little sonorant material is available. In particular, nuclear falling contours (H* L-%) tend to be truncated in L1 German while they are compressed in L1 English. Here we recorded 14 Bulgarian learners in their L2s German and English (within subjects, language order counterbalanced) when producing utterances in a statement context. The target word, a surname placed at the end of the utterance, differed in the available sonorant material (disyllable vs. monosyllables with long and short vowels). Our findings showed that Bulgarian speakers primarily truncate nuclear falling movements ((L+)H* L-%) in both L2s, suggesting transfer irrespective of the target strategy. However, our data show substantial inter- and intra-individual variation which we will discuss, along with factors that might explain this variation.

Intonational patterns of verbal irony: A cross-varietal study on two German regional accents

Sophia Fünfgeld, Angelika Braun and Katharina Zahner-Ritter

The present study investigates the intonational marking of irony in two regional accents of German, Moselle Franconian (Trier region) and Low Alemannic (Freiburg region). Results show that, irrespective of mode (ironic vs. sincere), differences across regions occur regarding the use of pitch accent types (overall more H* in the Trier region and more L*+H in the Freiburg region), and in the phonetic implementation of the pitch accents (tonal alignment in Freiburg tends to be later). To mark irony, speakers from both regions use accent position as a cue by placing an additional prominence in the prenuclear region (e.g., DAS sieht ja UMwerfend aus 'That looks stunning'). In nuclear position (e.g., umwerfend 'stunning'), the pitch range of the accentual movement is smaller in ironic as compared to sincere utterances (phonologically encoded by H* in ironic vs. L+H* in sincere utterances). This study thus provides initial insights into the interplay between regionally specified intonational patterns and the phonetic encoding of ironic attitude.

Processing prosodic boundaries in Dutch coordinated constructions

Rachida Ganga, Jorik Geutjes, Elanie van Niekerk, Victoria Reshetnikova and Aoju Chen

Across languages, major prosodic boundaries, such as intonational phrase (IP) boundaries, are typically signalled via final lengthening, pitch change, and pause. However, the relative weight of each cue in both production and perception is different across languages. Little is known about IP boundaries in Dutch. This study investigates cue-weighting in the processing of IP boundaries in Dutch by examining the effects of varying combinations of cues on the neurophysiological correlate of boundary processing, i.e., the Closure Positive Shift (CPS). Thirty native speakers of Dutch listened to a name sequence, connected by the coordinating conjunction en ('and'), i.e., Moni en Lilli en Manu. By leaving out one cue at a time, we have found that the CPS response was similar when listening to boundaries marked by all cues, boundaries marked by two cues missing pitch rise, and boundaries marked by two cues missing final lengthening, indicating that pitch rise and final lengthening have a relative small weight. In contrast, the CPS response was absent when listening to boundaries marked by two cues missing pause. These results indicate a crucial role for pauses in the adult processing of IP boundaries in Dutch coordinated constructions.

Arm movements increase acoustic markers of expiratory flow

Raphael Werner, Luc Selen and Wim Pouw

The gesture-speech physics theory suggests that there are biomechanical interactions of the voice with the whole body, driving speech to align fluctuations in loudness and F0 with upper-limb movement. This exploratory study offers a possible falsification of the gesture-speech physics theory, which would predict effects of upper-limb movement on voice as well as respiration. We therefore investigate co-movement expiration. Seventeen participants were asked to produce a continuous exhalation for several seconds. After 3s, they execute one of five within-subject movement conditions with their arm with and without a wrist weight (no movement, elbow flexion, elbow extension, internal arm rotation, external arm rotation). We analyzed the smoothed amplitude envelope of the acoustic signal in relation to arm movement. Compared to no movement, all four movements lead to higher positive peaks in the amplitude peaks, while weight did not influence the amplitude. We also found that across movement conditions, positive amplitude peaks are structurally timed relative to peaks in kinematics (speed, acceleration). We conclude that the reason why upper-limb movements affect voice loudness is still best understood through gesture-speech physics theory, where upper-limb movements affect the voice directly by modulating sub-glottal pressures. Multimodal prosody is therefore partly literally embodied.

Positional Effect in the Articulation and Acoustics of Stressed Vowels in Italian

Bowei Shao, Philipp Buech, Anne Hermes and Maria Giavazzi

Lexical stress may be signalled through a large number of acoustic parameters. In Italian, stress is realized through (1) longer duration, (2) more peripheral acoustic vowel shape, and (3) higher intensity. Moreover, duration has been shown to be sensitive to the position where stress occurs in the word: penultimate stressed syllables are longer than antepenultimate stressed syllables. Little is known however on other acoustic correlates and on articulatory correlates of this positional effect. Using EMA (AG501), we aim to investigate and to describe the interplay between the acoustic and articulatory parameters of this positional effect. The results show that (i) antepenul- timate stressed vowels are shorter than penultimate ones, (ii) antepenultimate and penultimate stressed vowels show a com- parable hyperarticulated pattern in tongue dorsum position and formant structure, (3) while antepenultimate stress' intensity in- creases in accordance with lip aperture, the penultimate stress' intensity peak occurs at the beginning of the vowel and is fol- lowed by a slope. This is not in accordance with the pattern of lip aperture. The findings are discussed within the hyperartic- ulation and the sonority expansion theories of prominence, and also within the framework of Articulatory Phonology.

Secondary prominence in Italian Southern varieties: the case of Cilentan

Giovanni Leo, Claudia Crocco, Mariapaola D'Imperio and Barbara Gili Fivela

In languages such as Italian, lexical stress can be cued by vowel duration. Stressed syllables can also attract F0 modulations in the shape of pitch accents. F0 modulations, however, have also been found to lend prominence to unstressed, pretonic syllables, both in Italian and other Romance languages. The present study investigates pretonic prominences on quadrisyllabic and trisyllabic words in Cilentan (spoken in Campania, South of Italy) with the aim of verifying whether these also entail a duration rearranging due to the lengthening of the pretonic stretch. We hence conducted both F0 and duration analyses on the pretonic stretch associated with the pretonic prominence. Results reveal that pretonic prominence presents higher scaling relative to nuclear accent H target and it is also associated with lengthening of the pretonic stretch, which appears mainly in trisyllables. We argue that the pretonic prominence functions as a secondary accentual prominence, mostly cued by F0, as opposed to the nuclear pitch accent occurring on the metrically stressed syllable and is cued by both F0 and duration.

Faster and smoother: Fluency in Chinese child-directed speech

Mengru Han, Lianghui Yang and Yan Gu

Child-directed speech (CDS) is often believed to have a slower speaking rate than adult-directed speech (ADS). This study examined the fluency between CDS and ADS as well as the individual differences in mothers' speaking rates. We annotated 2917 utterances in a corpus of Chinese ADS and CDS, where 19 mothers told the same story to their 24-month-old children and an adult. We coded and compared the fluency measures between ADS and CDS: speech rate (SR, including utterance-internal pauses), articulation rate (AR, excluding utterance-internal pauses), frequencies of silent pauses, filled pauses, repairs, and repetitions. We have three main findings: (1) CDS was generally more fluent than ADS, with fewer silent and filled pauses. (2) Contrary to common belief, only 7 out of the 19 participants showed a decreased SR and AR in CDS. (3) There were no significant differences in SR or AR between CDS and ADS when the utterance length was shorter than 4 syllables, whereas CDS was significantly faster than ADS when utterances were longer than 5 syllables. This suggests that Chinese CDS is not slower but instead faster than ADS. These findings highlight language-specific and individual variations in the temporal aspects of CDS.

The cognitive perspective on pre-planning sentence intonation: a cross-linguistic approach

Nele Ots

The study investigates linguistic factors (utterance length) and cognitive factors (scope of mental resources) influencing f0 in spontaneous speech. Specifically, it examines the connection between cognitive demands of language processing and length-dependent f0 raising, providing insights into the cognitive aspect of intonation planning. The experiment assessed the f0 difference between short and long utterances, spoken either in the presence or absence of a concurrent word recall task. The results reveal subtle yet consistent effect of utterance length on sentence intonation across both Estonian and German. Regardless of the language, f0 peaks tend to be higher in longer utterances, indicating effective cross-linguistic pre-planning of intonation for spontaneous speech production. With regard to cognitive factors, f0 peaks were low in Estonian but high in German in the presence of concurrent word recall task. Tentatively, the opposite effects of load on tonal scaling in two languages may have revealed the sensitivity of sentence intonation to different domains of working memory. The results, obtained through an innovative methodology, present novel findings, opening avenues for further exploration and understanding of the cognitive underpinnings of intonation planning.

Lexical encoding of Mandarin tones by L2 learners: A cross-linguistic study

Shuangshuang Hu

The present study, from a cross-linguistic perspective, investigated to what extent the native word and/or intonation prosody influence the lexical encoding of L2 Mandarin tones. Advanced Seoul Korean (SK) learners of Mandarin and advanced Russian learners of Mandarin were selected as participants in that SK and Russian differ in word prosodic system. Russian uses word stress while SK employs no word prosodic cues. A lexical decision task was applied where participants were required to judge whether the disyllabic words/nonwords they heard were real words. It was found that overall the advanced SK learners and the advanced Russian learners performed significantly worse than the natives, showing that their lexical encoding of Mandarin tones was not native-like. The advanced SK learners (with a rich inventory of pitch accents in the native intonation) significantly outperformed the advanced Russian learners (with comparably fewer intonational pitch accents) , suggesting that the native intonation prosody may better play a role in encoding Mandarin tones lexically than word prosody for non-tonal L2 learners. Furthermore, language-specific and language-general patterns were found in the lexical encoding of L2 Mandarin tonal contrasts, which were discussed in terms of perceptual models and the influence of the native prosody systems.

How to annotate prominences in schizophrenic speech? From manual to automatic processing

Simona Trillocco, Anne Lacheret-Dujour and Emanuela Cresti

This paper presents a prosodic analysis of schizophrenic speech in an Italian corpus based on prominence labeling. Four recordings from a clinical setting (about 40 minutes total) were orthographically and phonetically transcribed. The transcriptions and annotations were aligned on the speech signal at various levels, including phonemes, VtoV, words, speakers, and overlaps. The aim is to compare an annotation conducted on this corpus in the functional framework of the Language into Act Theory (L-AcT) to the automatic detection of prominences using the data-driven model of ANALOR. First, we present the perceptual methodology used to conduct the manual annotation of the corpus in line with the principles of L-AcT. Second, we describe the implementation of the prosodic model for automatic prominence detection and its application to our data. Finally, we compare the manual and automatic output and discuss the advantages and limits of automatic data-driven labeling. Initial results show a close correspondence between the two approaches because ANALOR predicts the prominences annotated manually. It can, therefore, be a robust tool for the automatic annotation of prominences in Italian pathological speech.

Can OpenAI’s TTS model convey information status using intonation like humans?

Na Hu, Jiseung Kim, Riccardo Orrico, Stella Gryllia and Amalia Arvaniti

Chatbots powered by Large Language Models (LLMs) such as OpenAI’s ChatGPT have demonstrated impressive capabilities in understanding and generating text and their potential applications in humanities research have been extensively explored. Recently, OpenAI launched its first Text-To-Speech (TTS) model, which has demonstrated the ability to convert text into highly realistic speech. This opens up various potential applications for prosodic research. However, before such applications are in place, a systematic evaluation is needed to determine the extent to which the synthesized speech resembles human speech in terms of prosody. This study aims to contribute to this endeavor by comparing how information status is conveyed by intonation in British English speech synthesized using OpenAI’s TTS model to the speech produced by native speakers of the same English variety. Through Functional Principal Component Analysis (FPCA) and statistical modelling, we found that OpenAI’s TTS model can generate F0 contours with various shapes. However, the F0 contours generated by OpenAI’s TTS model conveying information structure differ from those produced by the human speakers. This indicates that the speech generated by OpenAI’s TTS model may not be ready for use in prosody research, yet.

Adult readers signal metric and phrasing structure through acoustic variation in a Spanish children's book

Mara Breen, Sheyla Garcia, Genevieve Franck and Ahren Fitzroy

The current study investigated how adult speakers produce the children's book El Gato Ensombrerado, a Spanish translation of Dr. Seuss' The Cat in the Hat, which maintains the original book's consistent metric and rhyme scheme. Using linear mixed-effects regression, we assessed how metric and rhyme structure – in addition to lexical, syntactic, and semantic features – account for duration, intensity, and pitch variation on each syllable. Similar to previous English findings, we show that adult Spanish speakers consistently signal hierarchical metric structure: higher metric strength is cued by increasing duration, decreasing intensity, and decreasing pitch. In addition, speakers signal rhymes with intensity and pitch, though this effect is unlike that previously observed for English. These results demonstrate similarity in the realization of metric structure across languages, and further demonstrate the consistent cues to linguistic structure that child listeners receive through hearing children's books read aloud.

Investigating Mandarin Tone and Focus Prosody Production in Hong Kong Cantonese Speakers

Wenxi Fei and Yu-Yin Hsu

Second-language (L2) acquisition is influenced by the differences and similarities between a learner's first language (L1) and their target language. Because prior research has shown that Cantonese and Mandarin speakers employ different acoustic strategies to express focus in speech, this study examines how Hong Kong Cantonese (HKC) speakers, who are L2 Mandarin speakers, produce Mandarin focus prosody. Twenty HKC speakers completed a tone identification-production task with 72 monosyllabic Mandarin words used in the main experiment. Based on how well they performed this task, they were divided into two proficiency groups. The members of both groups then performed a speech-production task in which they answered pre-recorded wh-questions, focusing on either a numeral (ANUM), or a noun phrase (ANP). The results showed that the sampled HKC speakers did not use the typically observed HKC focus-marking strategy when producing Mandarin focus, but instead adopted a new acoustic strategy consisting of partial Cantonese focus marking strategy (lengthening) and some post-focus F0 compression, when speakers have a higher level of Mandarin proficiency. The two proficiency groups' acoustic approaches to expressing Mandarin focus differed from each other. These results represent an important contribution to our understanding of how HKC speakers perceive and produce Mandarin tone and prosody, and shed new light on L2 speech acquisition at both the suprasegmental and the sentential levels.

Duration as a prosodic marker of contextual factors in Mandarin positive polar questions

Xiaotong Xi, Siyu Zhou and Peng Li

This study investigated how Mandarin speakers use duration to mark contextual factors, namely epistemic bias, evidential bias, and the presence of antecedent, in their production of Mandarin positive polar questions, ending with either the particle -ba (PPQ-ba), -ma (PPQ-ma), or a prosodic pitch rise (PPQ-H%). Thirty native Mandarin speakers participated in a discourse completion task, producing sentences with a Subject-Verb-Object structure in response to different discourse scenarios. We designed 14 conditions, each encompassing four discourse prompts, to elicit sentences with all felicitous combinations of the contextual factors. Syllable duration was analyzed. Our findings revealed that contextual factors significantly influenced the syllable duration of PPQ-ba sentences but not that of PPQ-ma or PPQ-H% utterances. In PPQ-ba, the lengthening was localized to the verb in stimuli with negative evidence or a negative antecedent. We conclude that Mandarin speakers utilize duration to mark pragmatic functions in PPQs, but only to a limited extent, suggesting that duration is unlikely to be the primary prosodic cue in PPQ production.

The production of Mandarin neutral tone sequences by Hungarian learners

Kornélia Juhász, Katalin Mády and Huba Bartos

The present analysis aims to investigate how Hungarian learners produce neutral tone sequences in Mandarin declarative and yes/no interrogative sentences. In Mandarin, apart from the four full lexical tones, there is a fifth value: 'neutral tone', mostly carried by clitic-like particles. It is often assumed to be phonologically unspecified, its realization being highly dependent on the preceding full tone's coarticulatory effect, as well as on sentence type. The aim of this study is to shed light on how neutral tone, which is extremely plastic in its acoustic realization, behaves when concatenated into sequences of neutral tone syllables, in the production of Hungarian L2 learners, whose L1 is atonal. Interrogative and declarative sentence pairs were produced by two L2 learner groups (lower and upper intermediate) as well as by a native speaker group. F0-contours of utterance-final sequences of 3 to 4 syllables were compared by GAMMs: the 1st syllable determined the full tone context (four different values) followed by the sequence of neutral tones. Our results show that Mandarin native speakers discriminate interrogative and declarative f0-curves, which L2 learners failed to reproduce. However, they did manage to approximate natives in producing boundary tones irrespective of the value of the preceding tone.

Production of Mandarin tones by Japanese native speakers

Qi Wu

This study examined the acquisition of Mandarin Chinese tones, by native Japanese speakers with a focus on the T3 (low-dipping) tone. Previous research has shown that Japanese learners tend to confuse T3 with other tones, notably T2, but it remains to be investigated whether the tones neighboring T3 affect the accuracy of T3 articulation. In this study, the effects of syllable count, tone position, and adjacent tones on the accuracy of T3 production were investigated. The results showed that the accuracy of T3 in both disyllabic and trisyllabic words is significantly affected by its position in a word, with a higher error rate in the final position. Moreover, the tones preceding T3 significantly affect the accuracy of T3 production, and T3 has a higher error rate when the preceding tone was T1. Among the tonal combinations studied, T1T3 and T4T1T3 were found to be particularly challenging for Japanese learners. Acoustic analysis of F0 showed that Japanese speakers had a lower F0 at the beginning of T3 and a narrower F0 range than native Mandarin speakers. Additionally, the lowest F0 for the Japanese speakers occurred earlier in the syllable and was higher than that of the native Mandarin speakers.

WET: A Wav2vec 2.0-based Emotion Transfer Method for Improving the Expressiveness of Text-to-Speech

Wei Li, Nobuaki Minematsu and Daisuke Saito

Emotion transfer aims to extract the emotional state from a reference speech and use it for converting text to speech with the same emotional state. Previous methods usually use a reference encoder, which consists of convolution layers and recurrent neural networks, to extract the emotion features from the reference mel spectrogram. However, these methods fail to extract robust emotion features because they obtain features highly entangled with other components, like speaker identity. This may lead to an emotion mismatch between the reference speech and synthesized target speech. In this paper, we propose WET: a Wav2vec 2.0-based emotion transfer model which improves the capability of emotion feature extraction. By adding auxiliary classifiers, the extracted features can be highly related to the emotion category of the reference speech. In addition, we utilize relative attributes method to control the emotion intensity of synthesized speech which makes the results sound to be more natural. Finally, we evaluate our method using several objective and subjective metrics, and the experimental results show that our proposal can achieve higher accuracy of emotion transfer while ensuring the speech quality.

Production of Contrastive Focus in Mandarin-Speaking Children

Sichen Zhang, Aijun Li and Jun Gao

This study investigated Mandarin-speaking children's spontaneous speech production of contrastive focus in five-syllable SVO declaratives (constructed as Adjective1+Noun1+Verb+Adjective2+Noun2). Ten children aged 4;7-6;0 participated in a picture-guessing game with their parents, where they were prompted to produce the target structure with contrastive focus at different positions. The results showed a generally accurate placement of the focal accent, which improved with age. Nevertheless, 10.2% of the total contrastive foci were misplaced in wrong prosodic words, and 8.3% in correct prosodic words but on a wrong syllable. Additionally, 37.5% of contrastive foci on Noun2 were realized at incorrect positions, while very few on Verb were misplaced. Misaccenting Noun2 was the most common error, followed by Adjective1 and Adjective2. Acoustic analysis revealed children's awareness of using prosodic features including pitch range expansion, durational adjustment, and post-focus compression to encode contrastive focus. However, the performance varied across individuals and developed gradually. Furthermore, hesitation pauses were found in a third of the utterances. The findings suggest that the adult-like mastery of contrastive focus realization is still developing by age 6, which is closely related the complex interaction of lexical and phrase-level prosody in Mandarin Chinese.

Can Cantonese listeners identify the prosodic cues of sarcasm?

Chen Lan and Peggy Mok

This study investigated how prosodic features characterize Cantonese sarcasm and how these features help native listeners understand sarcastic meanings. 28 native Hong Kong Cantonese listeners listened to 50 sentences naturally produced by 6 native Cantonese speakers with a sarcastic attitude and with a sincere attitude. For each sentence the listeners rated whether they perceived the sentence as being produced with a very sincere (1) or very sarcastic (6) tone on a 6-point Likert scale. Acoustic analysis of the stimuli revealed that a slower speech rate, a lower mean F0, a lower mean amplitude, a narrower F0 range, a greater amplitude range, and a higher HNR value are all significant prosodic cues for Cantonese sarcasm although these cues may not be jointly present to deliver a sarcastic meaning. Listeners' ratings indicated that Cantonese listeners were able to discriminate sarcasm from sincerity based on the prosodic features. The combination of the prosodic cues used by the speakers influenced how well the listeners could perceive sarcasm. The more prosodic cues were utilized in a sarcastic speech, the easier it would be for the listeners to understand the implied sarcastic meaning.

Effects of part-of-speech, quantity and predictability on acoustic durations in Estonian spontaneous speech

Kaidi Lõo, Pärtel Lippus and Benjamin V. Tucker

Spontaneous speech is highly variable. Content words are produced with longer durations than function words; more predictable words are produced with shorter durations than less predictable words. These effects further interact with each other such that reduction due to predictability is more extreme for function words than for content words. However, these effects have mainly been investigated in English. The current study focuses on Estonian, a Finno-Ugric language with a more complex word class and phonemic quantity system than English. Our analyses indicate that content words (nouns and adjectives) are longer than function words (pronouns and conjunctions) in Estonian, even when controlling for predictors such as the number of phones, speech rate etc.

Whereas words are longer in the second and third quantities than in the first quantity, the effect is not the same for all parts-of-speech. Nouns and adjectives, are more affected by the quantity distinction than other parts-of-speech. The duration of content and function words decreases with increasing predictability.

In summary, we show that although the effects of part-of-speech and conditional probability in Estonian are overall similar to English, their exact influence is modulated by characteristic properties of the language in use, such as part-of-speech and quantity distributions.

Non-native lexical tones presented in low-high direction is beneficial for learning: Evidence from Cantonese rising and level tones

Ting Zhang, Mosi He and Bin Li

Learning Cantonese tones is challenging for Mandarin speakers. Various factors are reported to affect non-native tone perception, and they include both linguistic ones such as L1 status, pitch correlates and extra-linguistic ones like presentation order. Whether and how these factors function and interact in non-native tone learning are yet well understood. Cantonese rising (high-rising T25, low-rising T23) and level (mid-level T33, low-level T22) tone contrasts share similar contours but differ in pitch heights. The phonetic (dis)similarities turn these tone pairs most difficult for Mandarin listeners. This study focusing on these two contrasts designed a short-term auditory training to help Mandarin listeners improve their perceptual discrimination of Cantonese tones. The training used tone pairs in different within-pair presentation orders: low-high (T23-T25, T22-T33) or high-low (T25-T23, T33-T22). Tone type (rising vs level) and segment familiarity (familiar vs unfamiliar) were independent variables. Results indicated significant improvement in tone discrimination by Mandarin participants who were exposed to input presented in low-high order and that improvement among the rising tone pairs was notably greater. The beneficial effect of low-high presentation order suggested effectiveness of more discernible input in acquisition of non-native phonological contrasts.

Prosodic variability in marking remote past in African American English

Kristine Yu and Alessa Farinella

This paper explores variability in the fundamental frequency (f0) of utterances containing the remote past marker BIN in African American English, which has been described as having higher f0, intensity and duration relative to preceding material, and reduced f0 following, though with some interspeaker variability (Green et al. 2022). Here we re-analyze data from Green et al. (2022) to characterize the space of possible phonetic realizations of BIN utterances. We computed the 90th percentile f0 value in pre-/on-/post-BIN regions to create a 3-point "topline" f0 shape profile of the utterance (Cooper & Sorensen 1981) and performed time series clustering and principal components analysis (PCA). Two clusters were identified, one with higher f0 on BIN and lower f0 post-BIN, and one with lower f0 on BIN and higher f0 post-BIN. Results from PCA indicate speakers vary along two dimensions: one relating to pre-BIN f0 and one to post-BIN f0. Both dimensions were tied to f0 height on BIN, demonstrating the role that global aspects of the contour play in the variability. We show how the topline representation of f0 contour shape is robust to missing values and uncontrolled sentences and thus useful for naturalistic speech.

Vowel lengthening in L2 Italian and L2 French: a cue for focus marking?

Bianca Maria De Paolis, Federico Lo Iacono and Valentina De Iacovo

Our study investigates the role of vowel duration as a cue for focus marking in both L1 and L2 Italian and French. We aim to compare our data to highlight potential influences of the native language on L2 productions in the use of this cue.

The analysis involves task-elicited speech from 60 participants: 15 native Italian speakers, 15 native French speakers, 15 French learners of Italian (L2), and 15 Italian learners of French (L2). Participants produced the same target constituent under four information-structural conditions: background, broad focus, identification focus, and correction focus. Results reveal that the information-structural function significantly influences stressed vowel duration in native Italian, with identification-focus and correction-focus constituents bearing longer duration than background and broad focus. However, the same pattern does not hold in native French. Crucially, this distinction is mirrored in the production of non-native speakers. While Italian learners of L2 French, in fact, modulate duration based on the informational role of the constituent, French learners of Italian L2 do not. We discuss these findings in relation to previous findings on other prosodic and syntactic markers of focus. Results are commented in light of typological differences in discourse-prominence marking and theories of L2 prosody acquisition.

The contribution of speech timing, f0 change, and voice quality to perceived prosodic proficiency in L2: a cross-lingual perspective

Heini Kallio

Various studies have remarked the difficulty of the English stress and intonation systems to language learners. The current study takes a holistic approach to EFL prosody by investigating the contribution of speech timing, f0 change, and voice quality to the prediction of prosodic proficiency in EFL speakers with four different yet typologically close L1s. This study continues from an earlier one that suggests syllable duration to be the best feature for capturing relevant prominence characteristics of EFL speakers. However, the results varied depending on speaker L1. Acoustic parameters were analyzed from 216 read utterances produced by EFL speakers with either Czech, Slovak, Hungarian, or Polish as their L1. Generalized linear mixed models were used to inspect the effect of acoustic parameters and speaker L1 on prosodic proficiency assessed by 40 trained raters. The results show a significant effect of timing parameters on assessed EFL proficiency, but the effect is considerably stronger for Hungarian EFL learners than for the other groups. Parameters of f0 change, in turn, proved more significant for Hungarian and Polish than for Czech and Slovak EFL learners. Further research is suggested to scrutinize the L1 effect in producing EFL prosody.

What makes a conversational agent sound trustworthy? Exploring the role of acoustic-prosodic factors

Yuwen Yu and Sarah Ita Levitan

With advances in machine learning and speech technologies, conversational agents are becoming increasingly capable of engaging in human-like conversations. However, trust is crucial for effective communication and collaboration, and understanding the signals of trustworthy speech is essential for successful interactions. While researchers across disciplines have sought to discover the signals of trustworthy speech, mostly in human speech, in this paper, we explore the human perception of trustworthy synthesized speech. We present the results of a large-scale crowdsourced perception study, designed to investigate the acoustic-prosodic properties of trustworthy synthesized speech. Highly controlled parameters are manipulated to test the effects of acoustic-prosodic features including pitch, intensity, and speaking rate. To evaluate trust perception in contexts that require vulnerability and trust, a real-world application of emotional support dialogues is used. The findings of this work contribute valuable insights to improve the perceived trustworthiness of conversational agents.

Processing of Compound and Phrasal Prosody in (Canadian) English

Celeste Olson, Suzanne Curtin and Angeliki Athanasopoulou

English speaking adults use information about prosodic structure to make inferences about a sentence’s intended meaning and disambiguate, for example, between the compound word greenhouse and the phrase green house. Our study investigates how adult speakers of Canadian English use the prosodic information in novel compounds and related phrases to accurately identify the meaning of ambiguous sentences. We used eye-tracking methodology to investigate online processing during a forced-choice task. The target structures are modeled after compounds like coloring books and phrases like drinking milk. In the first experiment, participants heard the target compound or phrase only in the target sentence, and in the second experiment, participants heard both the compound and phrase before they heard the target sentence. Each target sentence was heard twice while participants saw the two drawings of each item/action. Participants in both experiments showed differences in their gaze patterns when hearing compound vs. phrasal prosody and exhibited a phrasal bias in the first experiment, but not in the second. These results indicate that adult speakers of English process prosodic information in sentence comprehension, but there are differences in how compound and phrasal prosody are used in this process.

Bilingual Production of Narrow Subject Focus in Japanese: Spelunking in Prosody

Onae Parker and Christine Shea

Languages differ in how they mark prosodic focus–whether syntactically/morphologically or phonetically/phonologically or both, which bears important implications for bilingual language acquisition. This study presents data on subject prosodic focus marking by L1 English/L2 Japanese, heritage Japanese speakers and L1 Japanese speakers. English primarily marks subject focus prosodically with stress and pitch fall, while Japanese also uses an obligatory postpositional particle, /-ga/. Fundamental frequency (F0) is used for subject focus in both languages, but intensity is used more consistently in English. Previous studies have examined L1 Japanese/L2 English production of subject focus, but there is little data looking at L1 English/L2 Japanese, and even less examining heritage Japanese speakers. Data was collected using a semi-spontaneous production task. Results show that i) L2 speakers used intensity more than the other groups; ii) HS speakers used a greater degree of F0 fall than the other groups, and the fall was greater when the /-ga/ focus particle was present; iii) the L1 speakers tended to employ greater F0 fall when the particle was absent. These results suggest that L2 learners transfer L1 prosodic cues to their L2. Heritage speakers, however, combine prosodic marking strategies from their dominant and non-dominant languages.

Tones do not disappear in singing: the duration of Mandarin tones in the music context

Qianyutong Zhang, Lei Zhu and Xiaoming Jiang

The primary acoustic cue of tones is F0, but secondary cues such as duration and intensity can also distinguish tones from one another. When pitch information is unavailable, do speakers utilize any of the secondary acoustic features to realize tones? This study situates Mandarin tones in a music context and explores tonal production patterns by speakers under different musical notes. The vowel duration of the four Mandarin tones in the context of normal speech and singing from sixteen native Mandarin speakers were analyzed. It was found that tones in the music context shows partially similar patterns with those in a normal speech context: syllables with T4 had significantly shorter vowel durations than those with other tones. However, the vowel duration of T1 increased, hence T3 was no longer the longest one. The results suggest that when pitch information is not available for the realization of tones, speakers may partially rely on vowel duration cues to express tonal contrasts, though the duration patterns of tones are affected by different communicative contexts.

Acoustic correlates of word stress in te reo Māori: Historical speakers

Kirsten Culhane

The current study is a preliminary investigation acoustic correlates of word stress in te reo Māori, the indigenous language of Aotearoa New Zealand. Māori is described as having a complex system of word stress, however, acoustic evidence has not yet been established. An additional consideration is that Māori has undergone various sound changes due to contact with English, and it is likely that the production and distribution of stress has also changed.

The aim of the current study is to determine if there was acoustic evidence for word stress in Māori prior to significant contact with English. This is done by using data from archival recordings of older Māori speakers born in the late 19th century. It investigates four potential acoustic correlates of stress: duration, spectral tilt, vowel quality and pitch. The results provide support for the production of word stress in Māori, as well as word positionality affects.

Prosody and gesture in the comprehension of pragmatic meanings: the case of children with Developmental Language Disorder

Albert Giberga, Alfonso Igualada, Nadia Ahufinger, Mari Aguilera, Ernesto Guerra and Nuria Esteve-Gibert

Prosodic cues facilitate children's understanding of pragmatic meanings. Multimodal prosody (i.e., combining prosody with body movements) provides enhancing cues to pragmatic comprehension, and could be beneficial for children with Developmental Language Disorder (DLD).

This study evaluated 45 Typically Developing children (TD) and 34 children with DLD (2 age groups: 5-7 and 8-10) in their ability to infer pragmatic meanings through prosody and gestures in a visual-world eye-tracking task. Pragmatic intent (interrogativity, indirect requests), and Experimental condition (prosodically-enhanced, multimodally-enhanced, non-enhanced) were presented within-participant.

Offline results revealed that prosody enhanced comprehension of the target meanings across groups ($\chi2$ =32.50; $p < 0.01$), that younger children with DLD are less accurate in general ($\chi2$ = 20.05; $p < 0.01$), and that multimodal cues especially help older children with DLD in complex meanings (indirect requests) ($\chi2$ = 4.1719, $p = 0.04$). Eye-tracking results showed faster and more accurate processing by older TD children over the other subgroups and no patterns of differences in the processing of multimodality.

Overall, we show how prosody and accompanying gestures can help children's pragmatic comprehension when structural linguistic abilities are compromised.

Voice transforms for affect control in Irish speech synthesis

Anna Maria Giovannini, Zihan Wang, Maria O'Reilly, Ailbhe Ní Chasaide and Christer Gobl

This paper reports on an experiment using voice transforms to alter the perceived affect in synthetic utterances of Irish, with a view to controlling affect in the spoken output of an Irish AAC device. The transforms were guided by prior experience and by voice source analyses of utterances by a male speaker with an angry, happy, sad, bored, relaxed and neutral voice. The neutral utterance was modified to incorporate stylized voice transforms targeting these affects. Modifications included global shifts affecting the entire utterance, local shifts affecting only accented syllables, and a combination of global and local changes. Stimuli targeting sad and happy included tempo changes and formant shifts were included for happy. Listeners' evaluations most positively identified the high activation affects happy and angry. Stimuli targeting sad were also effective, while those targeting bored and relaxed were not, although bored was positively associated with some of the sad-targeting stimuli. Results for low activations states are confounded by the fact that the neutral stimulus was to some degree biased towards bored, sad and relaxed affects. Of the three types of transforms, global, local and combined, the most effective appears to vary with the targeted affect.

The role of prosodic cues in the perception and expectation of sentence completion in structurally ambiguous sentences

Yujie Ji and Xiaoming Jiang

In spoken discourse, speakers strategically deploy prosodic cues to shape listeners' perceptions and expectations of sentence completion in structually ambiguous sentences. By asking 64 listeners to judge the degree of perceived sentence completion or to anticipate the ending point of the sentence in two experiments, this perceptual-acoustics study explored three aspects: (1) Acoustic distinctions at various levels of prosodic hierarchy between different ambiguity types; (2) The influence of prosodic cues at different levels on the perception of sentence completion and the classification of ambiguity types; (3) Which level of prosodic cues was employed by the listener to perceive and anticipate sentence completion. The findings revealed: (1) Structural ambiguity types impacted f0, intensity, and duration of boundary nouns; (2) The accuracy to classify structural types was higher in models with multi-level cues than those with single-level cues, and the highest accuracy was achieved with local cues of boundary nouns; (3) Local prosodic cues increased accuracy of perception and reduced RT for anticipation, while distal cues show no significant impact. These outcomes suggest that listeners rely on local parameters to determine spoken sentence completion, whereas speakers in oral communication are influenced by a "global" encoding mechanism.

Lexical stress perception by Trinidadian English listeners

Philipp Meer, Ronald Francis and Robert Fuchs

Research on lexical stress in postcolonial Englishes has focused on speech production. In the Caribbean island of Trinidad, production research indicates the existence of a lexical stress system in Trinidadian English (TrinE) and Trinidadian English Creole (TEC). However, as in postcolonial Englishes more generally, little is known about how Trinidadian listeners perceive stress.

Using a forced-choice identification task, the present study investigates the perception of lexical stress in disyllabic words by tertiary-educated speakers of TrinE (N = 46). Participants were presented with 21 truncated word pairs with segmentally identical first syllables but different lexical stress locations. One token had stress on the first, the other stress on the second syllable (e.g. car- in CARton vs. carTOON). Participants were asked to judge which of the words in a pair was the source of the current fragment. Results are analyzed using logistic mixed-effects models for trochaic (first syllable stressed) and iambic (second syllable stressed) words.

The findings suggest the existence of a lexical stress system in TrinE at the level of speech perception. They further indicate differences in stress perception across postcolonial varieties of English: there appears to be a greater sensitivity to word stress for TrinE listeners compared to IndE listeners.

Vocal emotion perception in adult cochlear implant users

Eleanor Harding, Etienne Gaudrain, Robert Harris, Barbara Tillmann, Bert Maat, Rolien Free and Deniz Başkent

Cochlear implants (CIs) provide hearing through electric stimulation to deaf or severely hard-of-hearing people. While speech intelligibility especially in quiet backgrounds is usually restored, perceiving prosodic changes in fundamental frequency (f0) that cue vocal emotions remains challenging. This is because the implant is typically unable to transduce fine-grained pitch information that reflects changes in voice f0. The current study tested 28 CI users with a vocal emotion task (emotions: happy, angry, and sad, expressed on a pseudospeech carrier). An ANOVA with sensitivity scores (d') showed an effect of Emotion, and post-hoc tests revealed that CI users categorized sad with more sensitivity than happy or angry. These results with vocal emotion align with patterns observed for a music emotion categorization task performed by the same participants as part of a larger test battery, suggesting that the emotion features best conveyed by the implant are consistent across speech and music.

Individual differences in pitch encoding and its use in phonological categorisation: integration from the bottom-up

Jiayin Gao and Justine Hui

This exploratory study addresses the relation between pitch encoding and the processing of phonetic details with an individual approach, from the bottom up to higher-level phonological processing. It reports on individual-based analyses from two experiments as a proof of concept for a larger project: (1) a preliminary report on the relation between JND (just-noticeable-differences) in pitch and duration, and the use of spectral and temporal cues in New Zealand English vowel perception, and (2) a re-analysis of the results from a perception study consisting of the use of VOT and pitch in the perception of [±voice] in French. By exploring individual differences in these two experiments, we advocate that it is important to consider individual variations on multiple levels of auditory and speech processing in an integrated manner. We further discuss a conceptual framework to dig into causal relations between auditory processing of spectral and temporal cues, language processing, and general cognitive mechanisms, based on large-scale participative research from typical and atypical listeners.

Delineating H* and L+H* in Southern British English

Jiseung Kim, Na Hu, Stella Gryllia, Riccardo Orrico and Amalia Arvaniti

In English, H* is said to encode new information and be realized as high pitch, while L+H* encodes degrees of contrastivity and realized as rising pitch. However, empirical evidence for this distinction is sparse, especially in Southern British English (SBE). To gain a better understanding, we examined 2,126 words with high and rising accents in an SBE corpus of unscripted speech. The accents were separately annotated for (i) f0 shape (high or rising) in PRAAT and (ii) pragmatic function (corrective, contrastive, and non-contrastive) based on written transcripts only. The data were modelled using Functional Principal Component Analysis and GAMMs. Phonetically H* and L+H* were distinct: H* was realized as a fall and L+H* as a rise-fall. However, these shapes did not neatly map onto pragmatic functions: corrective accents were likely to be L+H*s and were, thus, distinct from non-contrastive accents, which were likely to be H*s, but there was no phonetic difference between contrastive accents and non-contrastive accents and corrective accents, indicating that the mapping between phonetic form and pragmatic function was not one-to-one. By separating the shape- from the meaning-based annotation, the relation between the f0 shapes and the pragmatic functions of these accents is thus better understood.

Effects of coda consonants on preceding vowel F0

James Kirby, Rasmus Puggaard-Rode, Sireemas Maspong and Francesco Burroni

This paper documents the acoustic effects of onset and coda consonants on the F0 trajectory of the preceding vowel in the Austroasiatic language Eastern Khmu. In addition to an onset voicing contrast, Eastern Khmu permits syllables ending in sonorants, unreleased plosives, a voiceless glottal fricative, and a glottal stop. While glottal stop and fricative codas are frequently implicated in tonogenesis, their coarticulatory effects on the pitch of the preceding vowel are reported to vary. Here, we analyze data from 20 speakers of Eastern Khmu producing monosyllables in which both onset voicing and coda type were varied. Growth curve analysis indicates that, relative to sonorant-final syllables, the presence of a non-sonorant final raises F0 of the preceding vowel for syllables with both voiced and voiceless onsets. Onset F0 effects are also visible in syllables with voiceless onsets, regardless of coda type. Coda-induced variations in the shape of the F0 trajectory are detectable early in the F0 excursion, a conclusion validated by a neural network classifier. We propose that the co-intrinsic effects of codas can impose a laryngeal setting affecting the F0 trajectory of the entire preceding syllable. We discuss the implications of these findings for classical models of tonogenesis.

Priming Boundaries in Production: Data from French

Dorotea Bevivino, Marie Huygevelde, Barbara Hemforth and Giuseppina Turco

Perception studies have shown priming effects of intonational boundaries in disambiguating sentences. Recently, we found evidence for boundary priming in the production of relative clause attachment constructions in English. This study investigates the generalizability of priming effects of boundary location in production in another language, French, characterized by a different way of organizing phrasing from English. We measured word-and-pause durations at critical locations in relative clause constructions, when participants repeated prime sentences and when they produced new ambiguous target sentences. Primes were manipulated to either include a boundary (half early, half late) or not.

Our results showed that in French, as in English, boundary location in primes affected both the repetition of prime sentences and the production of new ambiguous targets. When repeating primes, participants produced the longest word-and-pause duration at the primed critical location. When producing targets, the default late boundary preference at the relative clause boundary was smoothed if the prime presented an early boundary. Our results replicate priming effects of boundary location in production previously observed in English. Taken together, these findings support the robustness of the claim that boundaries can be primed and affect sentence processing, suggesting a somewhat abstract representation of prosodic structure computed in early stages of production planning.

The influence of conversational context on lexical and prosodic aspects of backchannels and gaze behaviour

Malin Spaniol, Simon Wehrle, Alicia Janz, Kai Vogeley and Martine Grice

In face-to-face communication, the interplay between gaze behaviour and backchannels (vocal feedback) is essential for shaping conversational dynamics. This study explores the intricate relationship between these elements in dyadic face-to- face interactions across three conversational contexts: getting to know each other, a task-oriented Tangram dialogue, and a spontaneous discussion. By integrating quantitative gaze data with qualitative speech annotation, we investigate the nuanced role of backchannels in various settings.

Results reveal distinct patterns in lexical choice and intonation across contexts. While the Tangram task elicits predominantly rising tokens, free conversations exhibit unexpected intonational variation. Contrary to expectations, averted/directed gaze does not explain intonational differences.

Surprisingly, mutual gaze precedes only a fraction of backchannels (1% in task-oriented speech, 27% in free speech), challenging the notion that mutual gaze invites backchannels. Mutual gaze during backchanneling varies by dyad.

This study contributes insights into the interplay of gaze behaviour, lexical and prosodic aspects of backchannels. It reveals the considerable role of conversational context in determining behaviour in both dimensions. By delving into the intricacies of human interaction, we advance our understanding of the subtle cues shaping the dynamics of face-to-face communication and pave the way for future investigations into the multifaceted nature of dyadic conversations.

Givenness perception in declaratives vs. exclamatives

Heiko Seeliger and Sophie Repp

In intonation languages like German, there typically is an inverse relationship between prosodic and discourse prominence. Given referents, whose discourse prominence is high, are marked with low prosodic prominence while new referents, whose discourse prominence is low, are marked with high prosodic prominence, with gradual differences along the given-new scale. Recent production studies on non-assertive speech acts show systematic deviations from this inverse relationship. In German exclamations, given referents regularly are produced with prominent accents, and overall are prosodically as prominent as new referents. In this paper, we study the perception of prosodic prominence in sentence types expressing different speech acts. In a rating study testing exclamatives and declaratives, participants rated the discourse prominence of an object referent with different degrees of prosodic prominence (deaccentuation, H* and L+H* accents) on a given-new scale. The results show an inverse relation of prosodic and discourse prominence for declaratives and exclamatives but the differences between accentuation and deaccentuation are significantly smaller for exclamatives than for declaratives. Thus, there is some decoupling of prosodic prominence and discourse prominence in exclamatives but not to the same degree as has been observed in production. We discuss potential reasons for this difference.

Prosodic prominence in Greek: methodological and theoretical considerations

Riccardo Orrico, Stella Gryllia, Na Hu, Jiseung Kim and Amalia Arvaniti

A popular paradigm for studying prominence is Rapid Prosody Transcription (RPT), in which linguistically untrained participants listen to utterances and mark on a transcript words they perceive as prominent. RPT responses can be sensitive to the type of instructions used, such as whether participants are asked to select only the most prominent word or all the words they deem prominent. Here, we compare results from two RPT studies with the same Greek materials but using the two above-mentioned instruction types. Inter-rater agreement scores were similar across the two studies and yielded comparable overall results (e.g., words in focus were more likely to be selected in both). However, the relevance of certain criteria varied depending on task: in the multi-word task, duration, amplitude, and F0max predicted prominence, while in the single-word task F0max was not a predictor. These differences suggest that the tasks investigated here are not interchangeable, despite similarities. Most importantly, they can each lead to different interpretations of what constitutes prominence, suggesting that researchers need to be cautious when using them.

Exploring the accuracy of prosodic encodings in state-of-the-art text-to-speech models

Cedric Chan and Jianjing Kuang

Modern speech synthesis models have achieved increasingly human-like outputs, and have particularly been shown to be practically indistinguishable from natural speech at the phone- and word-tiers. Still, many text-to-speech (TTS) models have been observed to contain errors at the prosodic level. The most commonly employed measures of synthesized speech quality, such as mean opinion scores, lack linguistically meaningful information about the prosodic plausibility of speech. In this paper, we explore methods for evaluating the effectiveness of prosodic encodings in language models by cross-analyzing state-of-the-art TTS models with corpus data of natural speech. Through automatic signal processing and exploratory statistical analysis, we examine an array of prosodic and acoustic features related to prominence and phrasing, including pitch, duration, and intensity. Our analysis suggests that the most significant among these prosodic indicators of TTS naturalness rely on correct assignments of major intonational events and phrasal pausing. Based on these results, we propose several quantitative measures to capture the prosodic accuracy of speech inputs. These results have important implications for furthering a theoretical understanding of perceptual importance in speech prosody, and also help reveal limitations of the prosodic knowledge in current deep learning speech technologies.

Understanding individual differences in audiovisual child-directed language: The role of empathy and personality traits

Yanran Zhang and Yan Gu

This study explores individual differences in broadcasters' use of child-directed prosody and gesture, focusing on the role of empathy and the Big Five personality traits. Forty-two female future broadcasters simulated live broadcasts for both adults (ADB) and children (CDB) programmes. Prosodic and gestural analyses showed several key findings. First, openness negatively predicted speaking rate, while empathy positively predicted the rate of representational gestures. Mean intensity was positively predicted by empathy but negatively by agreeableness in CDB. Additionally, the saliency of pointing gestures was positively influenced by empathy and conscientiousness. Furthermore, participants varied in adjustments between programmes. Compared to ADB, in CDB, prosodically, higher empathy and neuroticism but lower extraversion predicted faster speech; higher empathy, extraversion and lower openness predicted higher pitch; and higher empathy and extraversion, along with lower openness and agreeableness, predicted higher intensity. Gesturally, higher-empathetic participants produced more salient pointing and beats, while more extroverted participants made more salient representational gestures in CDB. Notably, the frequency of child-directed representational gestures negatively correlated with neuroticism. The findings highlight the role of individual differences in tailoring audiovisual child-directed communication, with implications for broadcaster training.

From old to new to contrastive: Exploring prosodic marking of information structure in child and adult speakers of Albanian

Enkeleida Kapia and Felicitas Kleber

Responding to a call for action for more research in understanding how children acquire prosodic marking of information structure (IS) in free word order languages (Chen & Narasimhan, 2022), this study investigates acquisition of prosodic marking of topic (old), rheme (new) and kontrast (contrast) (Vallduvi & Vilkuna, 1998) in one such and hitherto understudied language – Albanian. Using a question/answer dialogue game, we elicited spontaneous production from five children aged six to seven years old and five adults. Regardless of age, prosody seems to play the primary role and word order a secondary one: both cohorts signal IS constructs mainly by means of different pitch accents within the canonical SVO order. The non-canonical order OVS was used very rarely to signal topic. Children up to the age of seven produce significantly less rise-fall patterns which in adults typically occur in kontrast and rheme conditions. The less transparent mapping between the prosodic form and function of rheme and contrast in Albanian may contribute to the late acquisition of this pitch accent, which suggests that the route and rate of acquiring IS differs from language to language, depending on their phonological and syntactic structure.

A Perceptual and Acoustic Evaluation of the Prosodic Speech Patterns of Young Autistic Adults

Shawn Nissen, Cassidy Gooch, Annika Henderson, Rachel Child, Rebecca Lunt and Garrett Cardon

This study is an examination of the prosodic speech patterns of young adult autistic individuals. A ten-minute conversational speech sample was recorded from 11 autistic adults between the ages of 18 and 26 years and a matching group of neurotypical individuals. Perceptual evaluations of the speech recordings were conducted using the Autism Diagnostic Observation Schedule (ADOS-2; Lord et al., 2012) prosody rating scale. In addition, pitch and intensity mean and variability measures were extracted from the speech samples using Praat acoustic analysis software (Boersma & Weenink, 2023). Acoustic measures were chronologically coded to examine how prosodic characteristics might change as the speaker becomes more familiar with the conversation partner over time. Results revealed significant prosodic differences between autistic and neurotypical individuals, and as a function of a speaker's identified gender. The ADOS-2 rating scale was found to be ineffective in accurately identifying autistic individuals, supporting the use of a more comprehensive perceptual rating scale. No prosodic changes were found for either neurological group across the degree of conversation partner familiarity. It is hoped that the findings of this study will lead to better assessment and treatment for autistic young adults, thereby improving their daily functioning and quality of life.

Morphology renders homophonous segments phonetically different: Word-final /s/ in German

Dominic Schmitz and Dinah Baer-Henney

"Recent research challenges established models of speech production by revealing unexpected phonetic differences in phonologically identical elements induced by morphological structure. While established models assume that morphology does not play a role in later production stages (Kiparsky, 1982; Roelofs & Ferreira, 2019), it has been shown that English word-final /s/ duration is longest in non-morphemic contexts, shorter with suffixes, and shortest in clitics (Plag et al., 2017; Schmitz et al., 2021). Subsequent research found that such differences are not only produced but also perceived by listeners and able to influence comprehension (Schmitz, 2022).

Recently, Baer-Henney & Schmitz (2023) investigated if German speakers could use subphonemic durational cues to acquire the morphological categories singular (short word-final /f/) and plural (long word-final /f/) in an artificial language. However, the study revealed subphonemic cues were insufficient.

Building on English findings, the present production study examines subphonemic durational differences in German word-final /s/. Preliminary results (20 speakers, 800 data points) show significant differences between non-morphemic and plural /s/ duration ($p<0.001$; Cohen's $d=0.3$). The findings challenge established models, suggest that Baer-Henney & Schmitz's (2023) null-results were due to the reversed direction of durational cues used, and indicate morphological influences on speech production extend beyond English."

Polish vowels in infant-directed and adult-directed speech: an investigation using an electromagnetic articulograph (EMA)

Katarzyna Klessa, Anita Lorenc and Łukasz Mik

In this paper we report on the results of examining selected features of Polish vowels in adult directed (ADS) and infant directed (IDS) speech, using Carstens AG501 electromagnetic articulograph (EMA). According to phonetic-acoustic research results, the characteristics of Polish IDS include e.g. higher durational variability, wider range of f0 changes, and higher F1/F2 ratio than ADS. Here, we provide new insights into the two speaking styles by discussing the specificity of the articulatory movements during speech production. The present EMA dataset is the first collection for Polish enabling such a comparison. The session scenario assumed recording of short utterances in two experimental conditions (ADS and IDS). The dataset comprises movement trajectories for 14 sensors (and referential sensors) in the x-axis (front-back), y-axis (left-right), z-axis (top-down). We also recorded audio signal and video image of all sessions. Thus, the dataset supports a combined description of the EMA results and phonetic-acoustic features in ADS and IDS. The statistically significant differences between ADS and IDS relate to e.g. the lip corner and lower lip sensor positions along the x-axis, as well as the shape of the tongue mass expressed by the differences in the tongue back position along the x-axis.

Real-Time Relations Between Prosodic Features of Infant-Directed Speech and Infant Attention at 3 Months

Yannan Hu, Mark Hasegawa-Johnson and Nancy McElwain

Infant-directed speech (IDS) guides infants' attention during caregiver-infant interactions. Prior studies have demonstrated infant preference for IDS over adult-directed speech and linked fundamental frequency (F0) and its variability to infant attention. However, we know little about how various prosodic features predict real-time infant attentional responses.

In this study, we assessed the IDS prosodic features that best predicted infant attentional engagement during mother-infant play. Seventy-nine infants and their English-speaking mothers were video-recorded during an 8-minute play session at 3 months. From each IDS utterance, we extracted 83 F0 features (e.g., mean, range) and 130 features capturing multiple prosodic aspects (e.g., energy, spectral) using openSMILE. To measure attention, infants' direction of gaze was coded.

The prosodic features were selected with regularized regression and refined using backward regression. We assessed model performance with cross-validations. Results showed that greater F0 variability and sharper F0 increases towards the end of the utterance predicted greater increases in the proportion of time infants gazed at the mother's face or interaction-related objects following IDS. Among other features, Mel-frequency cepstral coefficients and harmonic-to-noise ratio were selected, suggesting that maternal vocal timbre and breathiness may also contribute to the regulation of infants' attention in real time.

Acoustic-Prosodic Cues to Trust and Mistrust in Spanish and English Dialogues

Yuwen Yu and Sarah Ita Levitan

Trust between conversational partners is critical for effective communication and collaboration. While numerous studies have examined spoken cues to deceptive speech to understand how untrustworthy speech is produced and perceived, little work has studied the characteristics of trusting speech, i.e. cues that indicate whether a speaker trusts their conversational partner. This is crucial for monitoring a speaker's perception of their interlocutor, which has implications for conversational outcomes. In this work, we examine trusting speech in both human-human and human-machine dialogues. We study this phenomenon across native speakers of three languages (American English, Mandarin Chinese, and Argentine Spanish) in order to examine how one's native language affects their production of trusting and mistrusting speech. We identify several stable acoustic-prosodic signals of trusting speech across speakers of different native languages and identify some notable differences. This work sheds light on the nature of trusting speech across settings such as culture, language, and domain. We build predictive models of trusting speech using acoustic-prosodic features, in both within- and cross-cultural settings. We study the interpretability of these models and use those insights to improve classification performance.

Time-series f0 analysis of contrastive tones: the case of Livonian natural speech

Tuuli Tuisk and Nicolai Pharao

Livonian is a Finnic language that differentiates between two contrastive phonological word tones: the broken tone and the plain tone. The broken tone is similar to the Danish stød in some respects and is said to be part of the word tone systems of the phonologies of languages in the Baltic region. The most characteristic acoustic reflexes of words with stød are the pitch contour's rising-falling or falling shape and (occasional) creaky phonation. These characteristics have been documented in controlled speech (e.g., Teras & Tuisk 2009). This paper presents findings for the tonal contrast in Livonian utterances in natural speech elicited from ten native speakers. To analyze the f0 contours associated with the stød contrast, generalized additive models (GAMs) were used. Interestingly, the findings show that the only clear difference occurs in disyllabic words in the speech of men. Men have a clearly rising-falling f0 contour in disyllabic words with stød, but a level-falling contour in words without stød. Women have a falling contour both in words with and without stød, i.e., the contrast appears to be neutralized in spontaneous speech for this group. The results indicate accommodation to the surrounding majority language of Latvian.

Consonant f0 effects. A case study on Catalan

Bogdan Pricop and Eleanor Chodroff

Consonant f0 (CF0), the phenomenon in which the vowel onset f0 tends to be higher following a voiceless consonant than following a voiced consonant, has been commonly observed across the world's languages. The present study examined this effect in vowels following stop consonants of Catalan, using one of the largest language-specific datasets available in the Mozilla Common Voice Corpus, which after filtering, contained over 150 hours of validated speech data from over 1000 speakers. The study investigated the magnitude of CF0 in Catalan in initial and late portions of the vowel, the linguistic and social factors that modulate this effect (e.g., stop place of articulation, following segment voice, stress, utterance position, phonetic voicing, gender, and dialect), and the degree of speaker variability within the language. While the effect is small and demonstrates considerable inter-speaker variability, our results nevertheless confirm that the CF0 effect is robust in the phonetic realization of Catalan voiced and voiceless stops in the initial and late portions of the vowel. Moreover, the effect of phonetic voicing goes in the opposing direction to CF0, suggesting that the phonetic targets corresponding to f0 are controlled during phonetic realization.

Interaction between speech planning and prosodic structure in English

Jungyun Seo, Ruaridh Purse and Jelena Krivokapić

It is well-established that pauses and final lengthening are phonetic correlates of prosodic boundaries. Another line of research has established that speech planning is known to be reflected in speech production; for example, longer pause durations are observed as the amount of planning needed for the upcoming utterance increases. Combining these two lines of research, the present study examined how speech planning and prosodic structure interact in speech production, focusing on temporal properties. Two questions were addressed in an electromagnetic articulography study of American English. The first question examined the effect of an increase in planning load on gestural and pause duration at prosodic boundaries. The results showed that only pause duration, but not gestural duration, is longer when more planning is required for the upcoming material, implying that speakers use only pause duration for additional planning at prosodic boundaries. The second question tests the effect of an increase in planning load at word boundaries. Along with the insertion of a pause, lengthening of gestures before and after the pause was observed, indicating that speakers insert prosodic boundaries for planning. Implications for the role of prosodic boundaries in speech planning and the nature of prosodic representation are discussed.

Language Proficiency and F0 Entrainment: A Study of L2 English Imitation in Italian, French, and Slovak Speakers

Zheng Yuan, Štefan Beňuš and Alessandro D'Ausilio

This study explores F0 entrainment in second-language (L2) English speech imitation during an alternating reading task (ART). Participants with Italian, French, and Slovak native languages imitated English sentences, and their F0 (dis-)entrainment was quantified using the Dynamic Time Warping (DTW) distance between the parameterized F0 contours of the imitated sentence and those of the models. Results indicate a nuanced relationship between L2 English proficiency and entrainment: speakers with higher proficiency generally exhibit less entrainment in pitch variation and declination. However, within dyads, the more proficient speakers demonstrate a greater ability to mimic pitch range, leading to increased entrainment. This suggests that proficiency influences entrainment differently at individual and dyadic levels, highlighting the complex interplay between language skill and prosodic adaptation.

Task complexity and pausing behavior in L1 and L2 task-oriented dialogues

Lucia Mareková and Štefan Beňuš

Examining how cognitive demands and complexity in comparable L1 and L2 communicative tasks affect prosody in general, and interactional pausing behavior in particular, offers a fruitful testing ground for predictions of speech production models and may also yield useful pedagogical implications. Twelve pairs of undergraduates played a collaborative game of giving directions structured into three levels of gradually rising complexity in both their native Slovak and non-native English. All silent and filled pauses were identified and information of their distributions, durations, and co-occurrences was extracted. The results suggest minimal effects of complexity, or its interaction with language, on pausing behavior. This is in line with the Levelt´s (1989) speech production model, in which higher complexity levels place greater pressure on the conceptualization stage of speech production, which differs only slightly between L1 and L2 speech production.

The role of pitch accent in discourse comprehension and the markedness of Accent 2 in Central Swedish

Hatice Zora, Helena Bowin, Mattias Heldner, Tomas Riad and Peter Hagoort

In Swedish, words are associated with either of two pitch contours known as Accent 1 and Accent 2. Using a psychometric test, we investigated how listeners judge pitch accent violations while interpreting discourse. Forty native speakers of Central Swedish were presented with auditory dialogues, where test words were appropriately or inappropriately accented in a given context, and asked to judge the correctness of sentences containing the test words. Data indicated a statistically significant effect of wrong accent pattern on the correctness judgment. Both Accent 1 and Accent 2 violations interfered with the coherent interpretation of discourse and were judged as incorrect by the listeners. Moreover, there was a statistically significant difference in the perceived correctness between the accent patterns. Accent 2 violations led to a lower correctness score compared to Accent 1 violations, indicating that the listeners were more sensitive to pitch accent violations in Accent 2 words than in Accent 1 words. This result is in line with the notion that Accent 2 is marked and lexically represented in Central Swedish. Taken together, these findings indicate that listeners use both Accent 1 and Accent 2 to arrive at the correct interpretation of the linguistic input, while assigning varying degrees of relevance to them depending on their markedness.

Crosslinguistic transfer of alignment patterns: The timing of prenuclear rising accents in English-German bilinguals

Gwen McGuire, Kristin Smith, Jennifer Dailey-O'Cain and Anja Arnhold

This study examined crosslinguistic transfer of alignment patterns in four groups of English-German speakers. Given previous findings of crosslinguistic transfer, we hypothesized that characteristics of English accent timing would be apparent when non-native German speakers spoke German and vice versa. We evaluated data from 13 second-language (L2) German speakers and five heritage German speakers (all native speakers of Canadian English), six native German speakers living in Canada, and ten native German speakers living in Germany. All participants were recorded reading 15 German and 15 English sentences. The alignment of F0 minima and maxima was analyzed with linear mixed-effects models in 784 prenuclear rising accents.

Findings showed both the low (L) and high (H) tones tended to align later in English utterances than in German utterances. Additionally, native German speakers (in Germany) had significantly earlier L and H when speaking in German and English, compared to native Canadian English (L2 German) speakers. This suggests the transfer of a speaker's native language alignment pattern when speaking in their L2. The heritage German speakers showed similar alignment patterns to L2 speakers, whereas Germans in Canada showed an intermediate pattern between native German speakers in Germany and native Canadian English (L2 German) speakers.

Intonational and durational features of the Asturleonese substrate in Northwestern Peninsular Spanish

Víctor Bargiela and Paolo Roseano

The regions in Northwestern Spain where Asturleonese has historically been spoken, continue to manifest distinctive prosodic patterns that markedly deviate from the prosody of Castilian Spanish, typically associated with the prestige and standard variety of the language. This investigation aims to describe diachronic features influencing prosody in León, a region where only vestigial rests of the Asturleonese substrate have persisted since the early 20th century. To this end, we have compiled a corpus comprising interrogative sentences articulated by elderly informants from a rural area in León and by younger speakers in major cities of the region (León, Zamora, Salamanca, and Palencia). The corpus has been semiautomatically annotated using Sp_ToBI.

The prosodic patterns associated with absolute interrogative sentences across all informants exhibit falling intonational contours aligning with those observed in the utterances of Asturian speakers in Asturias, diverging from the rising patterns of Castilian Spanish. Moreover, a distinctive contour, linked to specific durational features, involving phonological lengthening of the final syllable, is discernible exclusively in the elderly speakers from the rural area. This observation could confirm the possibility that modality is encoded not solely through F0 movement but also duration as a durationally specified contour.

Prosodic and gestural marking of focus types in Catalan and German

Alina Gregori, Paula G. Sánchez-Ramón, Pilar Prieto and Frank Kügler

Prosody and gesture coordinate with each other in speech. This study investigates the contribution of prosody and gesture to the marking of focus types comparing Catalan and German. We hypothesize that multimodal prominence increases along increasing layers of pragmatic meaning in focus, namely background (no focus) < information < contrastive < corrective. To test this, we conducted a semi-spontaneous production experiment in Catalan and German, that systematically varied eliciting contexts for each focus type, in which participants (n per language = 15) interacted with a digital character. Target focused adjectives (791 items) were annotated for pitch accentuation and gesture presence, as well as degrees of perceived prosodic and gestural prominence.

Results suggest that while target adjectives are systematically produced with pitch accentuation across focus types in both languages, the number of head gestures tends to significantly increase with focus pragmatic strength in both languages. Crucially, a significant positive correlation was found between focus types and degrees of perceived prosodic and gestural prominence in both languages. In short, an increasing multimodal marking was observed across focus types, in terms of number of gestures produced and perceived prosodic and gestural prominence, which indicates an integrated behavior of prosody and gesture in speech.

Stability of prosodic performance over the lifespan: The (late) Queen's speech

Sam Hellmuth

Studies of individual sound change over the lifespan benefit from availability of regular broadcast recordings by individuals whose role keeps them in the public eye over multiple decades. Annual Christmas (and other) broadcasts of the late Queen Elizabeth II have served to disentangle the impact of linguistic sound change on vowel quality from the natural effects of aging. In this paper we show that, in contrast to the well-documented changes in segmental vowel quality, key prosodic features of the late Queen's broadcast speech in the public domain changed little over 70 years. Five speeches are analysed, ranging in broadcast date from 1947 to 2017. Each speech was segmented into intonation phrases based on auditory impression and each phrase coded for discourse structure. Values of maximum F0 per phrase show expected age-related decline over the lifespan, but linear regression models indicate minimal change across recordings in speech rate (sylls/sec), number of words and syllables per phrase, and duration of phrases and inter-phrase unfilled pauses; we ascribe the few observed differences to specific contextual factors (e.g. ill health or challenging content). Overall, we interpret this stability as a marker of individual performance style across the lifespan, within a restricted discourse genre.

Perception of Emotional Valence Projected by Prosody in Prefaces to Announcements of News

Emilie Marty, Caterina Petrone, James German and Roxane Bertrand

Prefaces (“I’m calling about your medical examination”) to news announcements ("They’re better than previously”) can serve important interactional functions such as preparing listeners for the emotional content of the upcoming news. We address whether, in French, listeners associate the valence of the upcoming news with a matching affective reaction depending on the prosody conveyed solely in the preface. For this, we used prefaces extracted from recordings of read voicemail-like announcements, presented in isolation (i.e., without the news). Prefaces varied in the emotional valence of the upcoming news: positive, negative, neutral. 110 participants (66 women, 42 men, 2 non-binary) were assigned to three experimental groups which varied in the pairing of two of three possible valences. They were instructed to choose among two-alternative forced-choices depending on their affective reaction. Results showed that only positive and neutral news elicited matching responses, and solely when they were contrasted with negative news. Additionally, women chose matching responses more often than men when positive and negative news were paired. No effects were found on reaction times. These findings suggests that prosodic variation within the preface leads listeners to associate the emotional valence of upcoming news with matching affective reaction.

Exploring the variations in disyllabic lexical tone sandhi in Xiangshan Chinese

Yibing Shi, Francis Nolan and Brechtje Post

Lexical tone sandhi in Northern Wu Chinese languages typically adopts a left-dominant mechanism, usually characterized as a tone extension pattern from the leftmost tone rightwards spanning the whole sandhi domain. By default, the underlying tone of the leftmost tone should be the only major factor conditioning the sandhi outputs in a lexical sandhi domain. However, variations still exist in some languages, the causes of which have not been studied extensively.

This study examines the categorical variations in disyllabic lexical tone sandhi in Xiangshan Chinese, an under-studied Northern Wu variety. A perceptual analysis of 106 disyllabic lexical compounds produced by 8 native Xiangshan speakers revealed 4 distinctive sandhi patterns. These patterns were cross-checked by independent k-means clustering based on acoustic data. Further analyses of the cluster distributions suggest that the historical tonal categories of the initial tones, and individual speakers, serve as the major sources of the variations. The results provide evidence supporting the reality and relevance of historical tonal categories in tone sandhi, and suggest distinct development paths for sandhi tones and citation tones. Inter-speaker variations largely explain the remaining variation and may reflect different stages of ongoing language change.

Prosodic characteristics of English-accented Swedish neural TTS

Christina Tånnander, Jim O'Regan, David House, Jens Edlund and Jonas Beskow

Neural text-to-speech synthesis (TTS) captures prosodic features strikingly well, notwithstanding the lack of prosodic labels in training or synthesis. We trained a voice on a single Swedish speaker reading in Swedish and English. The resulting TTS allows us to control the degree of English-accentedness in Swedish sentences. English-accented Swedish commonly exhibits well-known prosodic characteristics such as erroneous tonal accents and understated or missed durational differences.

TTS quality was verified in three ways. Automatic speech recognition resulted in low errors, verifying intelligibility. Automatic language classification had Swedish as the majority choice, while the likelihood of English increased with our targeted degree of English-accentedness. Finally, a rank of perceived English-accentedness acquired through pairwise comparisons by 20 human listeners demonstrated a strong correlation with the targeted English-accentedness.

We report on phonetic and prosodic analyses of the accented TTS. In addition to the anticipated segmental differences, the analyses revealed temporal and prominence-related variations coherent with Swedish spoken by English-speakers, such as missing Swedish stress patterns and overly reduced unstressed syllables. With this work, we aim to glean insights into speech prosody from the latent prosodic features of neural TTS models. In addition, it will help implement speech phenomena such as code switching in TTS.

Go local or go long: The relationship between dependency length and prosodic prominence in the production of Mandarin-speaking adults and children

Kexin Du, Li Zheng, Sergey Avrutin and Aoju Chen

While various sentence processing models predict that longer syntactic dependencies are more taxing to process due to the higher information content, few have investigated how dependency length influences the use of prosody, and how children acquire this capacity. This study tackles these two issues via comparing anaphoric dependencies produced by Mandarin-speaking adults and children. In the ambiguous Mandarin sentence: “Boris dreamt that Miffy painted zi-ji ’s face”, the anaphoric element ‘zi-ji’ can have two potential dependencies (allowed by grammar): (a) a local dependency where ‘zi-ji’ refers to Miffy; (b) a non-local dependency where ‘zi-ji’ refers to Boris. Using a picture-matching game, we elicited such sentences in both interpretations from Mandarin-speaking adults and 5-6 year-olds (20 per group). Linear-mixed-effect modeling of relevant prosodic parameters revealed that (1) the longer the dependency, the longer the duration of the anaphor ‘zi-ji’; (2) in the non-local dependency, the non-head syllable ‘zi’(falling tone) bears the prosodic prominence of longer duration and a lower minimum pitch; (3) despite exhibiting similar tendency, 5-6 year-olds’ prosodic production is not adult-like yet. This result is consistent with the Smooth Signal Redundancy Hypothesis, which postulates that prosodic prominence (duration in particular) coincides with the language units of higher information content.

Age-dependent intonational changes in child-directed speech

Daniil Kocharov and Okko Räsänen

The linguistic properties of child-directed speech (CDS) change over time as children get older and their language skills develop. The focus of this research is on prosodic changes of CDS within the earliest years of children's life, especially on the changes in melody. We analyzed mothers' speech from Providence corpus, a collection of longitudinal (bi-monthly) recordings of mother-child spontaneous speech interactions from six English-speaking children between 1.0–3.5 years of age (363 h of audio). Raw prosodic features were extracted from speech using OpenSMILE toolkit. Timing of prosodic events with respect to segmental content was estimated with automatic alignment of orthographic transcripts and the speech signals. Analyses of prosodic features in the data show that mothers' voice in CDS changes during the second and the third years of their children life, as the mean fundamental frequency lowers significantly, while the within-utterance fundamental frequency variability doesn't change.

Examining melodiousness in sarcasm: wiggliness, spaciousness, and contour clustering

Csilla Tatár, Jonathan Brennan, Jelena Krivokapić and Ezra Keshet

This paper compares ways of quantifying the phonetic correlates of sarcasm focusing on two novel measures, wiggliness and spaciousness [Wehrle, Cangemi, Krüger, & Grice (2018) Proceedings of AISV], neither of which has been examined in affective prosody before. We compare these to further F0 measures. In a production study, American English speakers (N=12) were recorded producing identically worded utterance pairs presented in contexts conducive to sarcasm and sincerity. Utterances were analyzed for wiggliness, spaciousness, F0 range, and the SD of F0 mean. The measures were entered into logistic regression models as predictors for sarcastic affect (wiggliness and spaciousness jointly, the others separately); model fit was evaluated with pseudo-R statistics. Results show that wiggliness and spaciousness together are comparable to F0 mean SD and F0 range in that they distinguish sarcasm from sincerity for many of the speakers (N=8). To compare the F0 contours in the two affective conditions, by-speaker contour clustering was performed [Kaland (2021)]. Preliminary results show the ability of wiggliness and spaciousness to capture differences between sarcastic and sincere F0 contours. There is variability in the F0-related phonetic correlates of speakers' sarcasm, but for many, it may be characterized by reduced wiggliness and spaciousness.

The effect of primary and rhythmic stress on onset consonant duration in Polish

Beata Lukaszewicz and Anna Lukaszewicz

Polish is known for having two levels of stress – penultimate primary stress and iterative rhythmic stress on odd-numbered syllables, forming a typologically rare binary system with internal lapses. Rhythmic stress has been reported to be cued by increased onset consonant duration relative to unstressed positions, both word-initially and word-medially. Little is known about the potential role of consonant duration as a phonetic correlate of primary stress. Hitherto available acoustic studies, relying mostly on vocalic parameters, often point to a strong dependence of those parameters on the focus position. In this study, we report acoustic results of an experiment designed to measure onset consonant (as well as vocalic) duration across the syllables of segmentally matched three- and four-syllable words, e.g. telefon [010] ‘telephone (nom. sg.)’, telefon+y [2010] (nom. pl.); 0 = unstressed, 1 = primary stress, 2 = rhythmic stress. In order to disentangle the potential effects of sentence-level prominence, the stimuli appeared in discourse-new, contrastive focus, and no focus contexts. Our results demonstrate significantly increased onset consonant duration in primarily/rhythmically stressed syllables relative to unstressed syllables, occurring to a different degree both in the presence and absence of focus.

Clustering approaches to dysarthria using spectral measures from the temporal envelope

Eugenia San Segundo Fernández, Jonathan Delgado and Lei He

Several clustering techniques were used for finding subgroups of speakers sharing common characteristics within a sample of 14 dysarthric speakers and 15 non-dysarthric speakers. Our classifying variables were five spectral measures computed from the temporal envelope of each of the four sentences read by the participants. The unsupervised k-means clustering algorithm showed that the optimal number of clusters in this dataset is two, with Cluster 1 matching almost exactly the dysarthric population and Cluster 2 the non-dysarthric population. As for the importance of each variable, a PCA analysis revealed that centroid, spread, rolloff and flatness contribute equally to the first component, and entropy contributes to the second component. Hierarchical agglomerative clustering further supported the separation into two main clusters (highlighting the relevance of these rhythmic measures to characterize dysarthria), but also allowed us to detect possible subgroups within each main speaker group.

Length affects the positioning of French attributive adjectives - Evidence from perception and production

Anna Pressler, Frank Kügler and Gerrit Kentner

This paper investigates the placement of a set of French attributive adjectives within noun phrases, characterized by flexible pre- and postnominal positioning without altering interpretation. We examine the effect of the prosodic factor length (in number of syllables), predicting a preference for short-before-long ordering. In addition, we explore prosodic cues inside the noun phrases. Two studies were conducted: a forced-choice task using written material and an elicited production task employing spoken material. The material manipulates the relative length of adjectives and nouns (longer, equally long or shorter adjectives) and their position (prenominal or postnominal). In the elicited production task participants combined two sentences (eliciting adjectives and nouns separately) to create adjective-noun pairs without being primed for one order.

Results suggest a preference for the short-before-long ordering in both perception and production. Additionally, the production data highlights that adjectives attain the highest F0 peak regardless of position, with different patterns for the F0 peaks of prenominal adjectives. These find-ings emphasize the impact of prosodic length on adjective placement in both perception and production and the alignment of F0 peaks within the noun phrases.

Investigating the Role of Prosody in Disambiguating Implicit Discourse Relations in Egyptian Arabic

Ahmed Ruby, Christian Hardmeier and Sara Stymne

We investigate whether prosody can help to disambiguate discourse relations. To address this question, we conducted a controlled experiment examining the impact of prosody in the absence of context, which is crucial for disambiguation. The aim was to determine whether specific prosodic features correlate with the disambiguation of implicit discourse relations. The dataset used in the experiment consisted of 22 pairs of examples, recorded by 21 native speakers of Egyptian Arabic. These examples are two-part sentences with an implicit discourse relation that can be ambiguously read as either causal or concessive, paired with two different preceding context sentences forcing either the causal or the concessive reading. We use linear mixed-effects models to analyze the impact of causal versus concessive discourse relations on prosodic features. We find that, relative to the causal relation, the concessive relation was produced with a longer pause duration between discourse segments, a wider F0 interquartile range for the second segment, and a lower last F0 max for the first segment. These differences are statistically significant, suggesting that speakers use prosody to distinguish between causal and concessive relations.

Analysis and Modeling of Self-Reported and Observer-Reported Personality Scores from Text and Speech

Soumik Dey, Guozhen An and Sarah Ita Levitan

Automatic personality detection from language has applications in diverse domains. Most previous efforts to automatically detect personality have aimed to predict either self-reported personality measures or personality labels provided by observers. It is unclear which kind of personality labels are preferable for personality recognition tasks or whether they are correlated with each other. In this work we aim to understand how self-reported and observer-reported measures of personality relate to each other.

We conduct a study of personality ratings by collecting personality judgments from external raters using a corpus of spontaneous speech from speakers with self-reported personality scores. We then proceed to build and compare predictive models of self-reported personality scores and observer-reported personality. Finally, we explore the effect of modality on observer-reported personality judgments, comparing judgments of audio stimuli with judgment of transcribed speech.

Word-final rhythmic prominence in Ukrainian

Beata Lukaszewicz and Janina Molczanow

As is well known, metrical prominence may coincide with boundary effects in word-final position. In this study, we report the results of an acoustic study of the metrical system of Ukrainian, designed to disentangle the potential domain-final lengthening from lexical and rhythmic stress in that language. The experiment is based on three-syllable words, differing in the position of lexical stress and rhythmic structure. We compare segmentally matched final syllables in triplets of words exhibiting three different metrical patters: [102], [010], [201]. The results point to statistically significant differences in vocalic duration, which depend on the stress level (rhythmic = 2, unstressed = 0, lexical = 1), and thus indicate that Ukrainian has word-final rhythmic prominence independent of boundary effects. We supplement those results with data from another acoustic study based on intrinsic (within-word) comparisons of vowel durations in the penultimate vs. final syllables. The comparison is based on three-syllable words lexically stressed on the first syllable, with penultimate and final vowels belonging to the same segmental category. The vowel in the final syllable (the rhythmic stress position) is significantly longer than the preceding vowel (the unstressed position). Our results corroborate the existence of word-final rhythmic stress in Ukrainian.

Pausing strategies in dialogue speech: the interlocutor factor

Tatiana Kachkovskaia and Daniil Kocharov

In everyday conversations, speakers tend to adapt their voices in order to fit the communication situation. One of the factors that seems to influence speech variability is the interlocutor factor: it has been shown that some linguistic features of our speech are dependent on who we are talking to. In our previous work, we observed interlocutor-related changes in speech tempo and frequency of non-speech events. In this research, we are presenting data on interlocutor-induced variability in within-turn pausing strategies, involving frequency of silent intervals, filled pauses and prosodic boundaries, as well as overall percentage of time taken by silent intervals in speech. The results were obtained using the large annotated speech corpus SibLing where each of the 20 core speakers participated in 5 dialogue settings: with a sibling, a friend, 2 strangers (a male and a female), and a person of elder age whose job requires leadership skills.

Methodological Influences on Word Stress Identification: Implications for Research and Teaching

Mahdi Duris, John Levis, Reza Neiriz and Alif Silpachai

Accurate word stress influences the intelligibility of L2 and L1 English speakers, making it important in areas such as language assessment and perception training. Accurate rating of word stress accuracy is assumed to be straightforward. However, reliably rating word stress in English is difficult because stress is signaled by four possible acoustic correlates (pitch, duration, intensity, and vowel quality), which are not always present in all spoken words. Multiple cues mean that direct judgments of word stress involve embedded decisions, leading to widely varied levels of agreement in published studies.

To investigate the influence of methodological decisions on word stress judgments, we employed two approaches to stress identification. Three phonetically-trained expert listeners rated stress placement by 10 Chinese L1 speakers of English who each read 100 multisyllabic English words (2-6 syllables), 1,000 words in total. In the first approach, raters identified whether each word was correctly stressed. In the second, listeners made a series of binary decisions about stress placement, syllable count, and vowel quality. Rater scores resulted in agreement levels among all three listeners of 41% (for approach one) to 91.6% (for primary stress placement in Approach 2), showing that ratings of word stress are sensitive to construct definitions.

Between dialect and standard: Segmental and prosodic differences in Zurich Swiss German speakers

Marieke Einfeldt, Anna Huggenberg and Bettina Braun

We investigate segmental and prosodic realizations across two varieties of a speaker (Zurich German and Swiss Standard German). Ten native speakers from the canton of Zurich read alternative questions of the type ‘Do you want <target> or X?’ in their dialect (Zurich German) and in their standard variety (Swiss Standard German). Ten Standard German speakers from Germany read the standard stimuli as control. We analyzed word-initial stops and pitch accent realization of prenuclear rising accents on disyllabic, trochaic target words (e.g., lenis: backen ‘to bake’ vs. fortis: packen ‘to pack’). The results showed that voice-onset-time (VOT) for fortis stops was the longest in Standard German, followed by Swiss Standard German, which in turn had longer VOT than Zurich German; VOTs for lenis stops did not differ across varieties. Swiss Standard German fortis stops were numerically closer to Standard German controls than Zurich German stops were. Intonation was modelled using general additive mixed models. The results showed that, within-participants, prenuclear accentual rises in Zurich German had a smaller f0 range than in Swiss Standard German. In comparison to Standard German Swiss Standard German was very similar in f0, but Zurich German differed for most parts of the target word.

Functional modeling of F0 variation across speakers and between phonological categories: Rising pitch accents in American English

Jennifer Cole, Jeremy Steffman and Aya Awwad

The Autosegmental-Metrical model of American English distinguishes three pitch accents with rising F0 trajectories (H*, L+H*, L*+H), differing in peak alignment and presence vs. absence of a low pitch marking the rise onset. Empirical studies report additional distinctions in the dynamics and scaling of the F0 rise, raising the question of which properties best capture variation among accents. We use functional principal components analysis (FPCA) to examine dynamic properties of accentual F0 trajectories in data from an intonation imitation experiment. F0 trajectories from 70 speakers producing rising accents on the phrase-final (nuclear) accented word were submitted to FPCA. The first three PCs account for 95% of variation in F0 trajectories and each shows significant differences between the three rising accents. Variation in PC1 primarily relates to differences in the overall F0 level of the trajectory, PC2 captures differences in rise shape (scooped vs. domed rise) and PC3 captures fine variation from a following Low phrase accent. Alignment distinctions are distributed across all three PCs. Examination of individual speakers shows all use PC1 and PC2 to some degree to distinguish rising accents, with no trading relations. Rises are variously implemented through level or shape distinctions, to varying degrees across individuals.

Development Of The Rhythmically Coordinated Duet Of A Bird Species (Southern Ground Hornbills, Bucorvus leadbeateri)

Sita M. ter Haar

Comparing human with non-human animal vocalizations can give insight in evolution and mechanisms of vocal behavior. This study addresses rhythm, a feature shared between speech prosody and music, in Southern Ground Hornbills (Bucorvus leadbeateri). Adults in this bird species sing a rhythmically coordinated duet, alternating between the male and the female. Vocalizations were recorded in two developmental and social settings in zoo living Southern Ground Hornbills and analyzed for their rhythmic patterns. Preliminary results indicate juvenile Ground Hornbills produce similar calls as adults, but do not produce the rhythmic duet yet. Moreover, upon introduction of a new adult male to an adult human-reared female, the birds did occasionally, but not always produce a rhythmic duet. These preliminary findings suggest that either learning and/or physiological changes during development and possibly pair bonding induce the rhythmic duet. Further research using more diverse settings and longitudinal recordings (including more adult controls), should disentangle these possible mechanisms. These findings are relevant for both evolution and mechanisms of rhythmic capacities, as well as animal welfare, particularly breeding programs to reintroduce Ground Hornbills to the wild. If development of the coordinated duet is in jeopardy, the breeding and reintroduction may be at risk as well.

The Adaptive Value of Mandarin Tones for Affective Iconicity

Tingting Zheng, Clara C. Levelt and Yiya Chen

Affective iconicity is an important mechanism for signaling emotions due to its adaptive significance. However, existing evidence predominantly focuses on emotional valence in Indo-European languages, leaving the more adaptation-related dimension, emotional arousal, relatively underexplored. Our study examined the adaptive significance of Mandarin tones in affective iconicity. We analyzed two Mandarin Chinese corpora datasets with arousal and valence ratings of bi-syllabic words. Hierarchical linear regression models were used to explore whether individual lexical tones associated with the first and second syllables of a disyllabic word predict arousal and valence ratings of the word. Results indicated that the valence of a word is predicted by the first tone, with negative words more likely to carry a falling tone than a rising tone. Lexical tones of both syllables predict the arousal rating of the word, with high-arousing words more likely to have a falling tone than a rising or low-dipping tone. These findings emphasize the importance of lexical tone pitch contour in predicting the affective iconicity of tone-carrying words. Our study extends the literature by showing that affective iconicity signals both dimensions of emotion-related adaptation (i.e., emotional arousal and valence) with evidence from a Sinitic lexical tone language.

Cue-weighting under focus: Predicting individual differences with autistic character traits

Monika Krizic, Daniel Pape and Gemma Repiso-Puigdelliura

Cue-weighting in perception and production demonstrates variation at the individual level, some of which can be attributed to different cognitive styles, such as autistic character traits. We examined the effects of autistic character traits on cue-weighting of prosodic focus, which is realized along multiple phonetic dimensions including f0, intensity, and duration. We report the results of 18 participants divided into two groups based on their Autism Spectrum Quotient (AQ) scores: a high AQ group (AQ > 124) and a low AQ group (AQ < 104). In a perception task, participants had to identify the noun in focus manipulated in f0, intensity, and duration, presented with natural-focus-production-range values (full condition) or below these values (half condition). In a production task, participants produced nouns in out-of-focus, broad focus, and narrow focus conditions. For perception, we found that high AQ participants differed significantly from low AQ participants by their perception of pitch (p<.019*) and intensity (p<.01**), but not duration (p<.158). For production, no significant effect of AQ level across any of the acoustic parameters was found. We conclude that perception is more sensitive to individual cognitive differences than production, and that, for perception, individuals with higher levels of autistic traits are better able to detect fine-grained differences in speech stimuli.

The role of prosodic structure in the planning of coordinated speech and manual gestures

Jelena Krivokapic, Mark Tiede, Martha Tyrone, Ruaridh Purse and Jungyun Seo

It is known that prosodic structure and co-speech gestures are temporally related to each other, but the exact nature of this coordination and how it arises is not understood. In this study, we test the hypothesis that the coordination is mediated by prosodic structure, with differing gestural types planned at different stages in the process of utterance generation, resulting in variation in the strength of coordination between speech and co-speech gesturing.

An electromagnetic articulometer (EMA) study was conducted, testing the effect of gesture type (beat gesture, deictic gesture) and utterance planning difficulty (easy, difficult) on the temporal properties of manual gestures and their coordination with co-produced speech gestures. The hypothesis predicts that beat and deictic gestures will be affected differentially by planning difficulty, reflecting how early in the speech planning process they are integrated with speech. Data from six speakers were collected.

Results provide qualified support for the hypothesis, reflecting in part individual speaker differences in planning strategies, and will be discussed in the context of their implications for models of speech planning.

PROTOSODY: A Semi-Automated Protocol for Experimental Prosody Research

Leônidas Silva Jr, Plinio Barbosa and João Marcelo Monte da Silva

This paper introduces Protosody, a semi-automated protocol developed using Python and Praat Scripting Language. This interdisciplinary approach combines computational methods with (foreign) language speech studies to enhance the extraction of prosodic-acoustic features in sound-transcription studies. The protocol operates on pairs of '.wav/.flac' audio files and '.vtt' files, transforming the linguistic data, numbers, punctuation, and orthographic diacritics into plaintext and retagging the files. These files are subsequently uploaded to a phonetic forced aligner. The protocol processes the returned forced-aligned TextGrids, converting phonemic-sized units into syllabic or higher speech units such as sentences or utterances, and defining the tonal range for each higher speech unit. The method extracts a comprehensive set of 74 features, encompassing four factors (Speaker_ID, Language-Dialect-Accent, Sex, Sentence-Utterance), 30 rhythm metrics, and 40 prosodic-acoustic features.

Pitch accents in Basque Spanish declarative utterances

Gorka Elordieta, Magdalena Romera and Asier Illaro

This paper is part of the project that analyses the intonational properties of the variety of Spanish in contact with Basque. The pitch accents and boundary tones of 360 declarative utterances from 12 speakers of Basque Spanish have been recorded and analyzed in conversational speech. In comparison, 210 utterances from 7 Spanish native speakers from Madrid have also been recorded and analyzed. In the abstract, we report the results from the analysis of 3,391 pitch accents. 72% of prenuclear accents and 63% of nuclear accents in the Basque Spanish data have peaks in the tonic syllable (L+H* and H*). In Madrid Spanish, on the other hand, the most common prenuclear pitch accent (48%) is a rising one with a peak on the posttonic syllable (L+<H*), and the most common nuclear accent is L* (51%). Interestingly, in the variety of Basque spoken in the area, recent work has shown that the most common prenuclear and nuclear pitch accent is L+H* (Ibarra 2023, Ezama 2023). Thus, a plausible hypothesis is that Basque L+H* is being transferred to Basque Spanish. Cf. Elordieta & Romera (2021, in press) and Romera & Elordieta (2020) for the presence of Basque intonational features in Basque Spanish interrogatives.

The prosody of Clefted Relatives: A new window into prosodic representations

Buhan Guo, Nino Grillo, Sven Mattys, Andrea Santi, Shayne Sloggett and Giuseppina Turco

The well-attested association between information structure and the acoustic properties of sentences can be captured by either assuming a direct mapping between semantics and acoustics or invoking the mediation of phonological processes operating on well-defined prosodic domains (indirect approaches). Although these two accounts' predictions typically converge, we identified an understudied contrast for which the two views make different predictions. Specifically, through 3 experiments (1 production, 2 comprehension), we tested the prosody of it-clefts containing string-identical Connected Clauses (-Who sang? -It was [the editor] [that sang]) or Relative Clauses (-Who called? -It was [the editor [that sang]] ([that called])) that have semantically focused elements of different structural sizes. Connected Clauses attach high in the structure and are given. Relative Clauses are assumed to convey background information, but here they are nested within the focused element and also in focus. Our production results showed a localized prominence on the rightmost stressable syllable of the Relative Clause, which is in line with indirect accounts. The comprehension studies further showed that i) clefted Relatives trigger garden-path effects in reading, but ii) garden-paths disappear when prosody is present. The studies support indirect accounts by employing more complicated structural configurations.

Measuring music and prosody: accounting for variation in non-native speech discrimination with working memory, specialized music skills, and music background

Adam Bramlett, Bianca Brown, Jocelyn Dueck and Seth Wiener

The dynamics of non-native speech perception remain poorly understood, especially in accounting for specialized skills/training. One such skill, musical ability, has been shown to positively impact sensitivity to speech sounds, yet how musical ability is operationalized and measured varies from study to study. Individuals' musical abilities vary in exposure-duration, skill type (e.g., voice, percussion), and skill-level. In the current study, we take an individual differences (n=44) approach to explore sensitivity in non-native speech discrimination of prosodic contrasts. We measure prosodic sensitivity, working memory, and three measures of musical ability: auditory-motor temporal integration [1], auditory discrimination [2, MET]), and musical sophistication [3, Goldsmiths-MSI]. We measured prosodic sensitivity using three AX discrimination tasks and signal detection measures (d'): Mandarin tone (primarily cued by pitch), Italian and Japanese (non-)geminates (primarily cued by duration). Results suggest music background, discrimination, and auditory-motor temporal integration capture related –yet divergent– aspects of music experience. Additionally, music sub- skills (e.g., pitch perception) have unequal contributions to non-native speech sensitivity across languages' respective linguistic cues (e.g., tone). Findings support models of non-native speech perception, which consider cognitive factors and auditory experience outside of language experience.

Effects of word order and embedded clause boundary on intonation in Tokyo Japanese

Shinichiro Ishihara and Joost van de Weijer

We investigate effects of word order, (embedded) clause boundaries as well as sentence parsing on intonation by analysing read speech elicited from native speakers of Japanese, a language with word order alternation.

All stimuli contain one dative phrase, which belongs to either the embedded or matrix clause.

a. NP1-nom [NP2-nom/dat NP3-dat/nom NP4-acc V1 Comp] V2
b. NP1-nom/dat NP2-dat/nom [NP3-nom NP4-acc V1 Comp] V2

We compared:
1) & 2) nominative–dative vs. dative–nominative order (in a and b, respectively) to test the word order effect;
3) the embedded vs. matrix dative phrase placed at NP2 to test the clause boundary effect.

Speakers read the stimuli twice, once instructed to read the sentence silently to parse it before reading it aloud for recording, and once to start reading aloud as soon as the sentence appears on the computer screen, to test the parsing effect. Words are either all lexically unaccented or accented to check the effect of the lexical pitch accent.

Results show significant effects on f0 by lexical pitch accents and parsing, but only partial effects by clause boundary, and no effects by word order.

Have four-year-olds mastered vowel reduction in English? An acoustic analysis of bilingual and monolingual child storytelling and comprehension

Paola Escudero and Weicong Li

Recent studies show that late Spanish-English bilingual adults tend to under-reduce or over-reduce some English vowels, while producing limited or no vowel reduction in their Spanish vowels. We examined whether children who acquire the two languages simultaneously show similar vowel reduction patterns. An acoustic analysis of English vowels produced by Spanish-English children during elicited story comprehension and retelling tasks was conducted following the same semi-automatic method used in previous studies on adult- and child-directed speech in English and other languages. Our analysis included 1136 vowel tokens from bilingual and monolingual children, with 714 tokens in stressed syllables. Bilingual four-year-olds had the same amount of vowel quality reduction in unstressed syllables as monolinguals and as adult vowel productions. For some vowels, bilingual children resemble adult vowel reduction even more closely than monolingual peers. However, divergence from monolingual vowel production was found in the F2 dimension. These results suggest that unlike late bilingual adults, simultaneous bilingual children resemble monolinguals in their English vowel reduction and acoustics, with some influence from their other language. The discussion includes explanations for the findings based on the Second Language Linguistic Perception (L2LP) model, which explains bilingual speech comprehension and production based on input and language inhibition.

Machine Learning Facilitated Investigations of Intonational Meaning: Prosodic Cues to Epistemic Shifts in American English Utterances

Nanette Veilleux, Stefanie Shattuck-Hufnagel, Sunwoo Jeong, Alejna Brugos and Byron Ahn

This work analyzes experimentally elicited speech to capture the relationship between prosody and semantic/pragmatic meanings. Production prompts were comicstrips where contexts were manipulated along axes prominently discussed in sem/prag literature. Participants were tasked with reading lines as the speaker would, uttering a target phrase communicating a proposition p (e.g., “only marble is available”) to a hearer who had epistemic authority on p. Prompts varied whether the speaker’s initial belief (prior bias) was confirmed (condition A: bias=p) or corrected (condition B: bias=¬p); this meaning difference was reinforced by response particles (A: “okay so” vs. B: “oh really”) preceding the target phrase.

Over 475 productions were annotated with phonologically-informed phonetic labels (PoLaR). To model many-to-many mappings between features (prosodic form) and classification (sem/prag meaning), Random Forests were designed on labels and derived measures (including f0 ranges, slopes, TCoG) from 299 recordings — classifying meaning with high accuracy (>85%). RFs identified condition-distinguishing prosodic cues in both response particle and target phrases, leading to questions of how/whether functionally-overlapping lexical content might affect prosodic realization. Moreover, RFs identified phrase-final f0 as important, leading to deeper edge-tone explorations. These highlight how explanatory ML models can help iteratively improve targeted analysis.

Tonal processing in Mandarin-speaking children with extensive cochlear implant experiences using an oddball paradigm

Ting-Syuan Wang, Pei-Tzu Liang, Chia-Lin Lee, Chen-Chi Wu, Tien-Chen Liu, Joshua Oon Soo Goh and Janice Fon

Tones are essential in differentiating word meanings in Mandarin and are predominantly realized through manipulation of fundamental frequency (F0). However, for children with cochlear implants (CIs), acquiring tones can be a challenge due to the limitation of CI devices in processing F0. Fortunately, research has shown that substantial CI experiences can potentially counteract the disadvantages CI children face in tonal acquisition, and help them achieve a production level similar to their normal-hearing counterparts. This study thus intends to investigate whether CI children could also perform equally well in tonal processing. Eight CI children were recruited five years after implantation, along with 16 chronological-age- and hearing-age-matched children with normal hearing. The passive oddball paradigm was used in an event-related potential (ERP) experiment, including a Tone 1/Tone 4 contrast and a Tone 2/Tone 3 contrast. Results showed that CI children had p-MMR and LDN in the Tone 1/Tone 4 contrast, and only p-MMR in the Tone 2/Tone 3 contrast. On the other hand, both groups of matched children with normal-hearing displayed LDN in Tone 1/Tone 4 and Tone 2/Tone 3 contrasts. This implies that lexical tonal processing in CI children might be less mature than that in their normal-hearing counterparts despite their near-normal performance in tonal production.

The kinematic profile of the Estonian ternary quantity distinction

Argyro Katsika, Eva Liina Asu, Matthew Gordon, Pärtel Lippus and Anton Malmi

This paper reports the results of an EMA study of the typologically unusual ternary quantity distinction of Estonian designed to examine the role of the metrical foot relative to the prosodic word in conditioning articulatory properties. Articulatory measures (formation duration, displacement, and peak velocity) of the three phonemic vowel lengths and the vowel's neighboring consonants, i.e., the onset consonant of the primary stressed syllable and the onset consonant of the first post-tonic syllable, were taken in words ranging from two to four syllables. Disyllabic and trisyllabic words both contain a single foot but differ in word length, while tetrasyllabic words consist of two disyllabic feet. Our results indicate a robust effect of phonemic quantity on the articulatory duration of the vowel. In addition, interactions were observed between word length and quantity, such that an asymmetry emerged along certain dimensions between two- and four-syllable words, on the one hand, and trisyllabic words, on the other hand. These findings support the view that the foot plays an important role in conditioning articulation in Estonian and that the articulatory signature of quantity does not funda- mentally differ from what is known about the articulation of stress in other languages.

Acoustic correlates of penultimate and final stress in Yami

Chun-Jan Young

Yami (Austronesian; Taiwan) has been impressionistically reported to exhibit default final-syllable stress for most content words, aside from the class of stative verbs which show penultimate stress. To empirically verify these claims, an experiment was conducted with 5 speakers contrasting the production of trisyllabic nouns (final stress) with stative verbs (penultimate stress). Duration, maximum and mean intensity, and maximum and mean F0 were measured for vowels in penultimate and final syllables (n=430). Linear mixed-effects models reveal that: for each word class, stative verbs showed significantly greater values in penultimate than final syllables for all variables, while final syllables of nouns were distinguished from penultimate syllables almost exclusively by higher maximum F0. When comparing syllables across word classes, penultimate syllables of stative verbs were greater in duration and F0 than nouns, but minimal difference was found between final syllables of stative verbs and nouns. Therefore, penultimate stress is robustly cued by duration, intensity, and F0, but stable correlates were not found for final stress, thereby challenging its validity. The higher maximum F0 of final (over penultimate) syllables in nouns could insinuate the presence of an accent, whose divergent behavior suggests that it may not be lexical, but rather phrasal, in nature.

Perceiving the social meanings of creaky voice in Mandarin Chinese

Yao Yao, Meixian Li, Shiyue Li and Charles B. Chang

While there is a growing literature on the social meanings of nonmodal voice qualities, most of the existing studies focus on English and use either naturally produced speech stimuli (which are hard to control acoustically) or a small set of fully synthesized stimuli. This paper reports a perceptual study of the social meanings of creaky voice in Mandarin Chinese, using a large set of resynthesized stimuli featuring 38 talkers (19F) and 6–10 pairs of sentences per talker that differed only in voice quality (creaky vs. modal). Sixty listeners (33F) answered 4 questions about the talker's demographic profile (age, gender, sexuality, education) and gave 19 ratings of personality traits (e.g., confident, professional, charismatic) and interactive potential (e.g., engagingness). Using factor analysis and mixed-effects modeling, our results showed that for male listeners, creaky voice significantly decreased the perceived warmth of male talkers but increased the perceived warmth of female talkers; creaky voice also led to more gender identification errors on female talkers by female listeners and made male talkers sound older. These findings point toward multifaceted social meanings of creaky voice in Mandarin, which extend beyond talker attractiveness and are closely linked to gender, both the talker's and the listener's.

Phonetic Realization of Focus in English by Taiwan Mandarin Speakers

Sherry Chien

This study examines how speakers of Taiwan Mandarin, a syllable-timed tone language, realize focus in L2 English. Taiwan Mandarin marks focus by increasing the pitch range and duration of the whole focused constituent, which can be polysyllabic. English, instead, marks focus via a pitch accent on the focused words' stressed syllable, which also lengthens. An interactive question-answering experiment was conducted. Two sets of initial-stressed English words featuring identical segments in the first syllable and varying word lengths were elicited phrase-medially under three focus conditions (contrastive, narrow, unfocused). Results show that stressed syllables undergo focus-related lengthening in monosyllabic words, but remain unaffected in polysyllabic words. Instead, focused polysyllabic words present lengthening on the first post-stress syllable. Meanwhile, F0 marks focus less robustly, being higher in focused than unfocused conditions only in polysyllabic words. Similarly to lengthening, the locus of the F0 effect is on the syllable following the stressed one. Neither duration nor mean F0 distinguishes between focus types. Results suggest that speakers of Taiwan Mandarin mark focus phonetically in L2 English, but the effects are not realized in the stressed syllable for polysyllabic words as is typical of English. These findings are discussed in terms of L1-L2 prosodic transfer.

Focus structure and articulatory strengthening in Seoul Korean

Jiyoung Jang and Argyro Katsika

Articulatory gestures under phrasal prominence undergo strengthening, becoming longer, larger, and faster. Limited research, mainly on head-prominence languages, shows that prominence-induced strengthening interacts with focus structure, increasing gradually across focus types. However, it is unclear whether focus structure is encoded in edge-prominence systems. Here, we turn to Seoul Korean, an edge-prominence language, in which the focused word starts an Accentual Phrase (AP) and exhibits prominence-induced strengthening, while the post-focal items are dephrased.

Analyses of kinematic duration, displacement, and velocity, examine degree of strengthening on focused AP-initial gestures and/or dephrasing on initial gestures in the first post-focal word. Results show that, in Korean, focused AP-initial strengthening reflects focus structure, although kinematic dimensions differ in the number of focus types they distinguish. Yet, the order of encoded types remains consistent and similar to that found in head-prominence languages. Post-focally, there is durational evidence of dephrasing only after contrastive focus and its reach is constrained by the number of intervening syllables. Instead, spill-over effects of focus are detected on the dimensions of displacement and velocity, indicating that focus-induced strengthening crosses word boundaries. These findings support the view that a hierarchy of prominence might emerge from the interface of prosodic structure with information structure.

Trisyllabic Tone Sandhi of Shaoxing Wu Chinese: Stress-conditioned, or Not?

Xinyi Wen, Yiya Chen and Lisa Lai-Shen Cheng

Stress has been proposed to condition tone sandhi patterns and domain formations across Chinese dialects (e.g., Duanmu, 2005). Empirical experimental data, however, are still needed to further attest this proposal. Shaoxing Wu Chinese, a Northern Wu dialect with complex tone sandhi patterns, can serve as a probe into the potential interaction between tone and stress. Previous studies suggest that in disyllabic words of Shaoxing Wu, both the initial and non-initial contours can be preserved depending on the tonal context. The current study took an experimental approach and examined the tonal realizations in trisyllabic compounds with 1+2 and 2+1 structures. Our results show that the general trisyllabic sandhi patterns are determined by the initial tones, with non-initial contours completely neutralized, which differs from the disyllabic pattern. In addition, 2+1 compounds consistently form one sandhi domain, while 1+2 compounds show variations in tokens with specific underlying tonal combinations (i.e., Falling + Falling + Rising). Such results are not predicted straightforwardly by a stress account of either the tonal changes or sandhi domain formation. Our findings call for a re-examination of the stress-based account for tone sandhi patterns in Chinese dialects.

Prosodic marking of information status in Chinese Sign Language

Hao Lin, Yi Jiang and Yan Gu

In spoken languages, new information is often produced with a longer word duration than given information. We investigate whether signers use the prosodic cue of duration to mark information status. Thirty-two deaf Chinese Sign Language (CSL) signers retold a story after watching a short cartoon clip. The data were glossed by a native CSL signer, with target references coded (e.g., different mentions of ‘bear’, ‘stone’). We examined whether there was any reduction in the references as a function of the information status. The results showed that first CSL signers mostly used nominals (24.0%), classifiers (42.1%), zero anaphora (30.0%), but hardly any pointings (3.9%). The nominals were both used in the first and subsequent mentions whereas classifiers, zero anaphora and pointings were almost always used in the non-first mentions. Furthermore, focusing on nominals, we compared sign durations over five mentions (M1=667.43ms; M2=426.09ms; M3=397.88ms; M4=440.03ms; M5=494.61ms). A regression analysis showed a significant linear and curvilinear relationship, indicating a gradient decrease in sign duration for the first three mentions and climbing up for the fourth and fifth mentions. In conclusion, CSL signers not only use linguistic devices to track references but also vary the duration of nominals to mark information status.

The German negative prefixes in- and un-: nasal place assimilation

Tina Bögel

The morphological and prosodic classification of the two German negative prefixes in- and un- has generally been based on the prefixes' behavior with regard to nasal place assimilation: in- is said to assimilate to the following plosive's place of articulation, while un- has been claimed to retain its alveolar nasal. The study presented in this paper investigated whether this long-standing claim can indeed be confirmed empirically via a production experiment that compared F2 trajectories of German in- and un-sequences followed by either an alveolar or a velar plosive: a) between words; b) at morpheme boundaries (i.e., as a prefix); c) within single morphemes. Results showed that un- does undergo nasal place assimilation as a prefix, in stark contrast to previous claims in the literature. Furthermore, a clear difference was found between the three contexts, with strongest assimilation patterns in the within-morpheme condition, weaker assimilation at the morpheme boundary, and no assimilation between words. This paper thus demonstrates the importance of empirical experimentation for the formulation of phonological generalizations.

Asymmetries of onset manner of articulation in the perception of tone register contrast in Wenzhou Wu Chinese

Weijun Zhang and Peggy Pik Ki Mok

The tone register contrast in Wu Chinese was developed from the consonant voicing contrast of the syllable onsets, and different contrasts in segmental features still co-occur with tone register in various prosodic positions. In general, the low register tones are related to voiced onsets and/or breathy vowels, while the high register tones occur in syllables with voiceless onsets and/or modal vowels. This study aims to investigate the perceptual cues of the tone register contrast in Wenzhou Wu Chinese between syllables with onsets of different manners of articulation. A generational change of the main perception cue from pitch to vowel phonation was observed, while the change was asynchronous between syllables with onsets of different manners of articulation. The perceptual role of pitch remained more robust in sonorant-onset syllables than in plosive-onset syllables. The perceptual importance of phonetic consonant voicing was found to be especially prominent in fricative-onset syllables, due to the equivalent phonetic realisation of tone register contrast in fricative-onset syllables, and possibly also cross-linguistic perceptual cue weighting strategies for perceiving consonant voicing.

The intonation of polar questions in Cypriot Arabic: prosodic contact in an endangered language

Mary Baltazani, Spyros Armostis and Elinor Payne

We analyse the intonation of polar questions (PQs) in Cypriot Arabic (CYA), a severely endangered peripheral variety of Arabic spoken by Cypriot Maronites, the intonation of which has not been formally examined before. We compare the patterns in CYA with those in Cypriot Greek (CYG) and those in Syrian Arabic (SYA) to search for possible differences with the latter, as a result of centuries of isolation, and similarities with the former, as a result of centuries-long contact on the island of Cyprus. We elicited PQs in CYG and CYA through a map task and analysed them combining quantitative modelling of intonational contours with conventional Autosegmental-Metrical tools of tune-text alignment. The results reveal that CYA questions are phonologically very similar to those in CYG, albeit with some differences in their fine phonetic detail. Polar questions in the Cypriot varieties of both languages have a L* nucleus followed by H-L% edge tones. This pattern is phonologically different from the pattern in SYA questions, which has been reported as a L* or a L*+H nucleus followed by a final rise. We discuss the implications of these findings for a theory of prosodic contact.

Effect of Sociolinguistic Variations on Rate and Rhythm of Hindi L2 Speech

Joyshree Chakraborty, Leena Dihingia, Priyankoo Sarmah and Rohit Sinha

The paper investigates the effect of Assamese L1 in native Assamese speakers speaking Hindi as an L2 in terms of speech rate and rhythm metrics in various sociolinguistic contexts. While Hindi is a syllable-timed language, the rhythm in Assamese varieties is reported to be akin to Japanese, which is a mora-timed language. We investigate spontaneous Hindi speech produced by 75 native Assamese speakers. A total of 4645 breath groups are analyzed for speech rate and rhythm using measures like syllable per second, segment per second, %V, nPVI, rPVI, VarCo-V, VarCo-C, DeltaV, and DeltaC. Assamese accented Hindi spontaneous speech data is compared with Hindi and Assamese read speech data and further grouped into age, gender, rural/urban area, and dialect. All these groups appear to show rhythmicity similar to Assamese read speech indicating L1 influence on L2 rhythm. Age and gender effects on rhythm measures are observed.