NVFW
Nederlandse Vereniging voor Fonetische Wetenschappen

# Dag van de Fonetiek

## ABSTRACTS

**8 november 2024**

Utrecht
Drift 21 – Sweelinckzaal

# Dag van de Fonetiek 2024

***8 november 2024***
***Sweelinckzaal, Drift 21, Utrecht***

ENTREE = GRATIS

| | | |
|---|---|---|
| 09:30 - 10:00 | **Inloop** | |
| 10:00 - 10:45 | **Keynote** | Turn taking in older adults and in persons with Parkinson's disease<br>*Esther Janse, Radboud Universiteit Nijmegen* |
| 10:45 - 11:00 | **Pauze met koffie & thee** | |
| 11:00 - 11:20 | Variatie en verandering in Friese vocaalbreking<br>*Cesko Voeten* | |
| 11:20 - 11:40 | Talker Familiarity as a Window into the Cognitive Architecture of Language<br>*Orhun Uluşahin, Hans Rutger Bosker, Antje S. Meyer & James M. McQueen* | |
| 11:40 - 12:00 | Individual variation in phonological repair strategies by Brazilian Portuguese-Japanese bilinguals<br>*Tim Laméris & Yōsuke Igarashi* | |
| 12:00 - 12:15 | **Algemene Leden Vergadering** | |
| 12:15 - 13:45 | **Lunchpauze** | |
| 13:45 - 14:05 | Intonation processing by Chinese speakers in imitation paradigms<br>*Wenwei Xu & Yiya Chen* | |
| 14:05 - 14:25 | Tonal contour clustering in Tongugbe Ewe: a preliminary investigation<br>*Man Yan Priscilla Lam & Yiya Chen* | |
| 14:25 - 14:45 | Lexical stress influences the perceived timing of beat gestures<br>*Chengjia Ye, James M. McQueen & Hans Rutger Bosker* | |
| 14:45 - 15:30 | **Posters & Koffie/thee** | |
| The Processing of Stress in End-to-End Automatic Speech Recognition Models – *Martijn Bentum, Louis ten Bosch & Tom Lentz* | | |
| The timing of an avatar's gestures differentially influences lexical stress perception in normal and simulated cochlear implant hearing conditions – *Matteo Maran, Roos Rossen, Renske Uilenreef & Hans Rutger Bosker* | | |
| Korean alveolar fricatives: Spectrographic evidence from running speech – *Patrik Hrabánek & Silke Hamann* | | |
| Bridging Boundaries: Combining Phonetic and Orthographic Information to Improve Automated Syllabification Performance – *Gus Lathouwers, Wieke Harmsen, Catia Cucchiarini & Helmer Strik* | | |
| Tonal adaptation of Japanese phonemic loanwords in Mandarin Chinese – *Jueyu Hou* | | |
| On the possibility of using pre-trained ASR-models to assess oral reading exams automatically – *Bram Groenhof, Wieke Harmsen & Helmer Strik* | | |
| 15:30 - 15:50 | "That's Fantastic!": The Prosodic Characteristics of Sarcasm in Childish Gambino's Speech, Rap, and Singing – *Valerie Querner & Steven Gilbers* | |
| 15:50 - 16:10 | Turn-taking in online interactions between people who do and do not stutter<br>*Lotte Eijk, Stefany Stankova & Sophie Meekings* | |
| 16:10 - 16:30 | A crosslinguistic study on prosodic characteristics in non or minimally verbal autism<br>*Laura Smorenburg, Jill Thorson, Aoju Chen & Wolfram Hinzen* | |
| 16:30 - 17:30 | **Borrel** | |

NVFW
Nederlandse Vereniging voor Fonetische Wetenschappen

**Turn-taking in older adults and in persons with Parkinson's disease**

*Esther Janse*
Radboud Universiteit

When we engage in a conversation, we often switch between the roles of speaker and listener, taking and yielding turns with our conversation partners. Language researchers have observed that time lags between consecutive turns tend to be very short or absent, leading to fluent turn transitions. Such fluent turn-taking must mean that 'next talkers' already prepare their turn while listening. At the same time, most turn-taking research, like any behavioural research, has based its claims on results obtained with student populations. One such claim is that listeners start planning their responses to questions as soon as they can. A second observation based on research with students is that 'next talkers' know or predict *when* to jump in, based on e.g., accurate perception of prosodic information signaling turn-finality of one's interlocutor. Where preparing one's turn while listening and knowing exactly when to jump in may be relatively easy for students, I will focus on turn-taking fluency in two populations in these two aspects of turn taking may be more difficult: healthy older adults, as well as people with Parkinson's Disease (PD). PD is best known for its effects on the limbs (tremor), but PD also affects speech acoustics and prosody, word finding, and cognitive control. Speed of information processing is often claimed to be lower in older adults (as compared to younger adults), and more particularly in those with PD. Decreased processing speed could be hypothesized to slow down responses to questions across the board, and could also compromise dual-tasking in conversation. Age-related hearing loss and Parkinson-related prosody perception problems may hamper accurate prediction of turn finality. In this talk, I will present the two turn-taking studies I am currently running, and present preliminary findings of one of them.

# Variatie en verandering in Friese vocaalbreking

*Cesko C. Voeten*[1,2]

[1] Universiteit van Amsterdam, Amsterdam Center for Language and Communication
[2] Fryske Akademy

Een klassieke en nog altijd onbegrepen variabele in het Fries is *breking* (Tiersma 1978), waarbij *ingliding* middenvocalen veranderen in stijgende diftongen (zodoende alterneert een woord als *stien* [stiən] “steen” met *stienen* [stjınə] “stenen”). De precieze condities waarin breking wel en niet optreedt zijn synchroon opaak; er zijn weliswaar uitvoerige descriptief adequate beschrijvingen (bijv. Van der Meer 1985), maar die zijn synchroon weinig plausibel (door bijv. veelvuldig gebruik van uiterst specifieke woordklassen waar geen externe synchrone evidentie voor is). Het hedendaagse begrip van breking gaat dan ook uit van synchrone *listing*, d.w.z. sprekers leren “simpelweg” van alle individuele woorden in het lexicon of zij wel of niet breken. O.a. als gevolg van (Arndt-Lappe & Ernestus 2020) de daaruitvolgende hoge synchrone psycholinguïstische belasting van breking (Arndt-Lappe & Ernestus 2020), is breking momenteel aan variatie én verandering onderhevig (cf. het zeer vergelijkbare fenomeen van lexicaal geconditioneerde split-[æ] in verscheidene Amerikaanse steden, die die conditionering nu aan het vereenvoudigen zijn naar een eenvoudige fonologische contextregel; Sneller 2018). De precieze vorm van deze variatie en verandering is echter sinds 1985 (De Graaf) niet fonetisch gemeten. In het NWO-project ‘The processing of language change – the case of Frisian vowel breaking’ wordt de hedendaagse fonetiek van breking onderzocht, én de psycholinguïstische verwerking daarvan. Hiertoe wordt begonnen met fonetische studie naar het Boarnsterhim-corpus van twee generaties van gesproken Fries, dat handmatig zal worden opgelijnd. Daarna wordt een attitude-experiment uitgevoerd naar de evaluatie van breking, en een MMN-experiment naar de cognitieve verwerking ervan.

**Referenties**


de Graaf, T. (1985). Phonetic aspects of the Frisian vowel system. *North-Western European language evolution* 5(1), 23-40.

Arndt-Lappe, S., & Ernestus, M. (2020). Morpho-phonological alternations: The role of lexical storage. In V. Pirrelli, I. Plag, & W. U. Dressler (Eds.), *Trends in Linguistics: Studies and Monographs: Vol. 337. Word knowledge and word usage: A cross-disciplinary guide to the mental lexicon* (pp. 191-227). Mouton de Gruyter.

van der Meer, G. (1985). *Frisian breaking: Aspects of the origin and development of a sound change*. PhD dissertation, Rijksuniversiteit Groningen.

Sneller, B. (2018). *Mechanisms of phonological change*. PhD dissertation, University of Pennsylvania.

Tiersma, P. M. (1978). Bidirectional leveling as evidence for relational rules. *Lingua* 45, 65-77.

Tiersma, P. M. (1979). Breaking in West Frisian: a historical and synchronic approach. *Utrecht Working papers in Linguistics* 8, 1-41.

Tiersma, P. M. (1979). Aspects of the phonology of Frisian based on the language of Grou. *Meidielingen fan de stúdzjerjochting Frysk oan de Frije Universiteit yn Amsterdam* 4.

Tiersma, P. M. (1980). *The lexicon in phonological theory*. PhD dissertation, Indiana University.

Tiersma, P. M. (1982). Local and general markedness. *Language* 58, 832-849.

Tiersma, P. M. (1983). The nature of phonological representation: evidence from breaking in Frisian. *Journal of Linguistics* 19, 59-78.

# Talker Familiarity as a Window into the Cognitive Architecture of Language

Orhun Uluşahin[1], Hans Rutger Bosker[2,1], Antje S. Meyer[1,2], James M. McQueen[2,1]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, NL

[2] Donders Institute for Brain, Cognition and Behaviour, Nijmegen, NL

While "a link" between language perception and production is axiomatic, theories differ extensively on what this link might resemble. One crucial question in this debate is whether these two modes of language use shared representations. Talker representations, formed exclusively through perception, present a unique opportunity to study the link between perception and production. We present three experiments that investigate the potential role of talker representations in speech production, as manifested in their impact on phonetic alignment to voice fundamental frequency (F0). In Experiment 1, female native Dutch speakers (N=32) performed a baseline reading task, followed by a synchronous speech task where they were instructed to synchronize (temporally) to a pre-recorded model talker. The model talker's voice was pitch-shifted to have high or low F0, and half of the sample performed the task in each F0 condition. We found that participants' F0 values tended to align with that of the model talker. In Experiment 2, we added an exposure task between reading and synchronous speech, during which participants (N=32) were familiarized with the same model talker at high or low F0. In the synchronous speech task, they heard the model talker at the familiar F0 condition. Finally, in Experiment 3, we reversed the F0 manipulation between exposure and test. Thus, participants (N=32) were familiarized with high or low F0 per their group, but heard the opposite F0 condition during synchronous speech. Cross-experiment analyses revealed that while congruent and conflicting talker information both diminish the effect size of alignment, only conflicting talker information reduces the probability of alignment. These results indicate that talker information acquired exclusively through perception is used during production.

# Individual variation in phonological repair strategies by Brazilian Portuguese-Japanese bilinguals

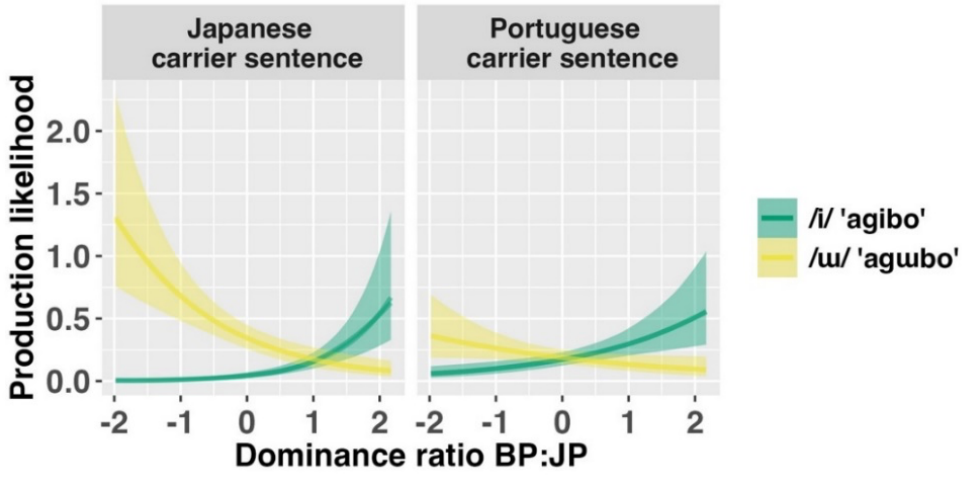
*Tim Laméris[1] & Yōsuke Igarashi[2]*

[1] Leiden University, [2]National Institute for Japanese Language and Linguistics

Japan is home to a Brazilian diaspora of approximately 200,000 individuals who are bilingual in Brazilian Portuguese (BP) and Japanese. I examine individual variation in phonological repair strategies in this bilingual community. Both Japanese and BP have a phonological repair strategy that involves vowel epenthesis in illegal consonant clusters and codas (Guevara Rukoz, 2021). Whereas Japanese typically inserts /ɯ/, BP inserts /i/. The English word 'laptop', for example, is produced as /ɾapːɯtopːɯ/ in Japanese and as /lɛpitɔpi/ in BP. The questions here are i) whether bilinguals apply these repair strategies separately in each language, or whether one language's strategy is applied cross-linguistically; and ii) whether language dominance influences repair strategies.

Speech data was analysed from 22 adult BP-Japanese simultaneous bilinguals in Japan. Language dominance (reflecting language proficiency and daily exposure) was calculated using the *LHQ3* (Li et al., 2020). Participants completed a multisyllable concatenation task to elicit vowel epenthesis in illegal consonant clusters. They heard a stimulus, e.g. /ag/, followed by a 500 ms pause and a second stimulus, e.g. /bo/, and were asked to produce the resulting nonce word in a Japanese or Portuguese carrier sentence. Acoustic analyses determined the presence and type of epenthetic vowel. **Figure 1** shows that speakers were likely to apply each language's strategy separately, producing /agɯbo/ in the Japanese sentence and /agibo/ in the BP sentence. However, individuals who were more dominant in BP were likely to apply the BP strategy, even in the Japanese sentence. This suggests that factors like language dominance have cross-linguistic effects on quite implicit aspects of speech production in highly proficient simultaneous bilinguals.

**Figure 1.** *Estimated count likelihood of /i/ or /ɯ/ insertion against dominance ratio.*




## References


Guevara Rukoz, A. (2021). *Decoding perceptual vowel epenthesis: Experiments & modelling* [Université Paris sciences et lettres]. https://theses.hal.science/tel-03288523

Li, P., Zhang, F., Yu, A., & Zhao, X. (2020). Language History Questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition*, *23*(5), 938-944. https://doi.org/10.1017/S1366728918001153

# Intonation processing by Chinese speakers in imitation paradigms

*Wenwei Xu*[1], *Yiya Chen*[1, 2]
[1]Leiden University Centre for Linguistics (LUCL)
[2]Leiden Institute for Brain and Cognition (LIBC)
w.xu@hum.leidenuniv.nl, yiya.chen@hum.leidenuniv.nl

In this study, we report data elicited with an imitation task on the processing of intonation contours by Chinese speakers. Two variant paradigms of this task have been used in the literature to investigate the imitation of (non-)native intonational contrasts. The immediate paradigm elicits imitation immediately after the stimuli are presented [e.g., 1, 2, 3]. In contrast, the delayed paradigm requires speakers to imitate after a 2-3 seconds' delay [4, 5]. The delayed paradigm has been argued to reflect phonological processing rather than phonetic (echoic) memory based on Baddeley's working memory model [6, 7, 8]. A direct comparison between the two paradigms, however, is not available in the literature.

20 Standard Chinese speakers with Mandarin and Wu dialectal backgrounds participated in the study and imitated pseudo sentences with nine synthesized intonation contours akin to typical intonation events in West Germanic languages. In the first block, they were asked to provide an immediate response, and in the second, a delayed response. Chinese participants were predicted to show more native language interference in the delayed block due to phonological processing [e.g., 5, 9, 10].

Fine-grained analyses of imitated F0 contours were performed using Generalized Additive Mixed Models (GAMMs). The results showed that participants could generally distinguish between all contours within each block. Moreover, significant differences in contour shape were found between immediate vs. delayed imitations for most of the contours. Visualization of the differences between the two variant paradigms suggest that for some contours, deviations in the delayed block are indeed more aligned with Standard Chinese intonation patterns; no patterns contradictory to the predictions were detectable. Jointly, our results lend evidence for phonological processing in delayed imitations.

## References


[1] Pierrehumbert, J. B., & Steele, S. A. (1989). Categories of tonal alignment in English. *Phonetica, 46*(4), 181–196. http://doi.org/10.1159/000261842

[2] Dilley, L. C., & Heffner, C. C. (2013). The role of f0 alignment in distinguishing intonation categories: Evidence from American English. *Journal of Speech Sciences, 3*(1), 3–67. http://doi.org/10.20396/joss.v3i1.15039

[3] Cole, J., Steffman, J., Shattuck-Hufnagel, S., & Tilsen, S. (2023). Hierarchical distinctions in the production and perception of nuclear tunes in American English. *Laboratory Phonology, 14*(1). https://doi.org/10.16995/labphon.9437

[4] Zahner-Ritter, K., Einfeldt, M., Wochner, D., James, A., Dehé, N., & Braun, B. (2022). Three kinds of rising-falling contours in German wh-questions: Evidence from form and function. *Frontiers and Communication, 7*. https://doi.org/10.3389/fcomm.2022.838955

[5] Zahner-Ritter, K., Zhao, T., Einfeldt, M., & Braun, B. (2022). How experience with tone in the native language affects the L2 acquisition of pitch accents. *Frontiers in Psychology, 13*, 903879. https://doi.org/10.3389/fpsyg.2022.903879

[6] Baddeley, A. D. (1986).Working Memory. Oxford: Oxford University Press.

[7] Baddeley A. D. (2003). Working memory and language: an overview. Journal of communication disorders, 36(3), 189–208. https://doi.org/10.1016/s0021-9924(03)00019-4

[8] Baddeley, A. D., & Hitch, G. J. (1974). Working Memory. In G. A. Bower (Ed.), Recent Advances in Learning and Motivation (Vol. 8, pp. 47-89). New York: Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

[9] Braun, B., Kochanski, G., Grabe, E., & Rosner, B. S. (2006). Evidence for attractors in English intonation. *The Journal of the Acoustical Society of America, 119*(6), 4006–4015. http://doi.org/10.1121/1.2195267

[10] Laméris, T. J., Li, K. K., & Post, B. (2023). Phonetic and Phono-Lexical Accuracy of Non-Native Tone Production by English-L1 and Mandarin-L1 Speakers. *Language and Speech, 66*(4), 974-1006. https://doi.org/10.1177/00238309221143719

# Tonal contour clustering in Tongugbe Ewe: a preliminary investigation

*Man Yan Priscilla Lam[1,2], Yiya Chen[1,2]*

[1] Leiden University Centre for Linguistics; [2] Leiden Institute for Brain and Cognition

In the initial stages of describing a tone system, it is a common practice to elicit controlled productions (of minimal pairs) and transcribe them based on auditory impressions. However, this method often involves few speakers and may be subject to the researcher's language experiences and auditory bias. Considering these limitations, *Contour Clustering* (Kaland 2023) potentially serves as a complementary method with a data-driven, automatic approach.

This study combines these two methods to investigate the tonal system and acoustic tonal space of Tongugbe, a lesser-studied dialect of Ewe (Niger-Congo; West Africa). Like other dialects, Tongugbe has three level tones: High, Mid, Low. Intriguingly, Tongugbe reportedly has two realizations of Mid, differing in height and duration (Kpoglu 2019; 2020).

We elicited 982 Ewe nouns produced by 26 Tongugbe speakers (local variety: Mepegbe), which yielded 1,412 f0 contours extracted from the rhyme vowel of each syllable. The wordlist consisted of a balanced set of three level tones. The tonal transcriptions are based on the literature and consultation of a native speaker-linguist, mainly using the impressionistic auditory description method. The contours were subjected to clustering analysis using the Contour Clustering application (Kaland 2023).

Preliminary results show four clusters of surface f0 contours (Figure 1). Cluster 2 and 4 may be the surface realization of High and Low tone syllables respectively, while Cluster 1 and 3 are possibly different realizations of the Mid tone, as proposed by Kpoglu (2019; 2020). A closer inspection of the results is planned to compare the tonal classifications based on the wordlist and the clustering analysis. With a combined methodological approach, results of this exploratory investigation will provide further insights on the Tongugbe Ewe tone system, thereby contributing to our understanding of the mapping between phonological tones and their phonetic realization.

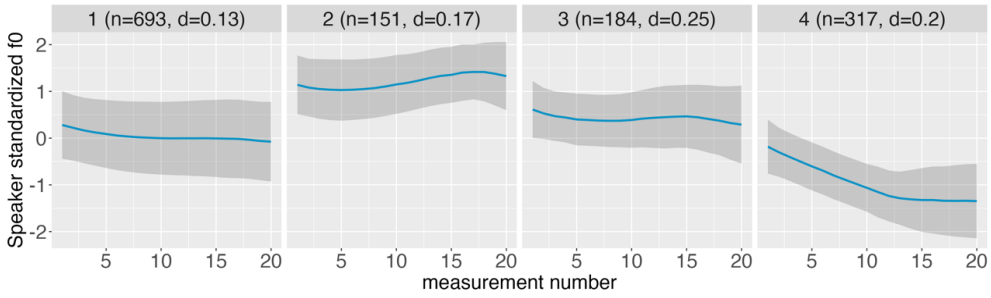



**Figure 1.** Visualization of the f0 contours obtained from cluster analysis with 4 clusters assumed. The four panels each represents a cluster.

## References

Kaland, Constantijn. 2023. Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours. *Journal of the International Phonetic Association* 53(1). 159–188. DOI: https://doi.org/10.1017/S0025100321000049

Kpoglu, Promise Dodzi. 2019. *Possessive constructions in Tongugbe, an Ewe dialect*. LOT dissertation.

Kpoglu, Promise Dodzi. 2020. The mid tone in Tongugbe, an Ewe dialect. In van der Wal, Jenneke & Smits, Heleen & Petrollino, Sara & Nyst, Victoria & Kossmann, Maarten (eds.), *Essays on African languages and linguistics : in honour of Maarten Mous*, 495–508. African Studies Centre Leiden (ASCL).

# Lexical stress influences the perceived timing of beat gestures

*Chengjia Ye* [1], *James M. McQueen* [1,2], *Hans Rutger Bosker* [1,2]

[1] Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands, [2] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Language is multimodal: for example, speech is often accompanied by hand gestures. Although beat gestures do not convey meaning themselves, they frequently co-occur with prosodic prominence. Thus, they can indicate stress in a word and hence influence spoken-word recognition. However, little is known about the reverse relationship. The current study investigated whether lexical stress has an effect on the perceived timing of hand beats. We used videos in which a disyllabic word, embedded in a carrier sentence (Exp 1) or in isolation (Exp 2), was coupled with an up-and-down hand beat, while varying their degrees of asynchrony. Using a novel beat timing estimation task, Experiment 1 revealed that gestures were estimated to occur closer in time to the pitch peak in a stressed syllable than their actual timing, hence a reduced temporal difference between gestures and stress. Experiment 2, a canonical 2AFC task, further demonstrated that listeners tended to perceive a gesture, falling midway between two syllables, on the syllable receiving stronger cues to stress than the other. The temporal attraction effect of stress on perceived gestural timing was greater when gestural timing was most ambiguous, and was driven by *f*0 and intensity. This study provides new evidence for auditory influences on visual perception, supporting bidirectionality in audiovisual interaction between speech-related signals that occur in everyday face-to-face communication.

# The Processing of Stress in End-to-End Automatic Speech Recognition Models

*Martijn Bentum[1], Louis ten Bosch[1], Tom Lentz[2]*
[1]Centre for Language Studies, Radboud University Nijmegen, [2]Tilburg University

Listeners use lexical stress to facilitate word recognition and speech segmentation. However, classical automatic speech recognition (ASR) models did not typically incorporate lexical stress in their recognition process. In contrast, end-to-end ASR models are trained in an unsupervised manner and may use the information carried by lexical stress.

The present study shows that Wav2vec 2.0 (an end-to-end ASR model) is indeed sensitive to lexical stress, and that this sensitivity is not a mere reflection of acoustic correlates of stress. Diagnostic classifiers of the convolutional neural network (CNN) output of the Wav2vec 2.0 model reveal vowel-specific stress representations, that perform on par with acoustic features. Stress classifiers trained on transformer layers of the Wav2vec 2 model outperform classifiers based on acoustic correlates, but degrade when context is removed, showing that later layers of the model take the relative nature of stress into account.

Results obtained by testing a lexical stress classifier on vowels it is not trained on, show that stress processing in the Wav2vec 2 model is to some extent abstract, i.e., the classifier does not simply detect a set of stressed vowel representations but rather, their common denominator.

# The timing of an avatar's gestures differentially influences lexical stress perception in normal and simulated cochlear implant hearing conditions

*Matteo Maran [1], Roos Rossen [1], Renske Uilenreef [1], Hans Rutger Bosker [1]*

[1] Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

Cochlear implants (CI) alter the transmission of sound, degrading fundamental frequency information, disrupting the perception of lexical stress. Providing additional (even artificially generated) visual cues could perhaps support CI users in speech perception. Recent studies showed that lexical stress perception is influenced by beat gestures: the same acoustic token (e.g., "*content*") is more likely to be perceived with a strong-weak (SW) stress pattern (e.g., the noun "*CONtent*") when the beat falls on its first syllable, and with a weak-strong (WS) stress pattern (e.g., the adjective "*conTENT*") when the beat falls on its second syllable ("manual McGurk effect"). The present study investigated whether beat gestures made by an avatar are beneficial when listening to vocoded speech, simulating CI conditions. Normal-hearing participants watched videos of a gesturing avatar, while a member of a disyllabic Dutch minimal pair was produced with a clear SW, a clear WS, or an ambiguous stress pattern, in normal or 8-channel tone-vocoded speech. The avatar's beat gesture fell on the first or second syllable of the word. With normal speech, beat gestures biased lexical stress perception mostly in the clear SW and ambiguous stress patterns. With vocoded speech, the manual McGurk effect was observed in the clear WS and ambiguous stress patterns. Since WS is an infrequent stress pattern in Dutch, the avatar's beat gestures appear to modulate lexical stress perception in conjunction with internal models of frequency of occurrence and the clarity of the speech signal. The present findings could inspire the development of avatars gesturing to the prosody of an interlocutor to support speech recognition in CI users.

# Korean alveolar fricatives: Spectrographic evidence from running speech

*Patrik Hrabánek & Silke Hamann*

University of Amsterdam

Seoul Korean features an uncommon distinction between the non-fortis fricative /s$^{h}$/ and the fortis fricative /s*/, both voiceless in word-initial positions. Previous studies show that /s$^{h}$/ is typically followed by aspiration, while /s*/ exhibits glottalization in about 50% of cases. Additionally, they differ in acoustic properties such as center of gravity (CoG), frication duration, and F1 values and duration of the following vowel.

The present study adresses the differences in the two fricatives using data from the *Seoul Corpus*, a corpus of running speech, as opposed to previous production studies which only considered single words or syllables. We also include younger speakers than in previous research, which is important due to a perceptual shift in those born after 1960.

As some /s$^{h}$/ tokens were surprisingly found to be phonetically voiced, we conducted two separate analyses. The first focused on 653 phonetically voiceless /s$^{h}$/ and /s*/ tokens, finding that 87% of /s$^{h}$/ tokens were aspirated, while 58% of /s*/ tokens were glottalized. The two fricatives significantly differed in frication duration, CoG, and vowel duration, with the frication duration effect being more pronounced for male speakers. The second analysis examined 151 phonetically voiced /s$^{h}$/ tokens, none of which showed aspiration. These tokens had significantly shorter frication duration, lower CoG (similar to values of /z/ or /ʒ/), and a slightly lower F0 of the following vowel compared to the phonetically voiceless tokens.

Challenges with distinguishing frication noise from aspiration noise based on spectrogram and waveform reading and the intentional exclusion of VOT measurements are discussed.

# Bridging Boundaries: Combining Phonetic and Orthographic Information to Improve Automated Syllabification Performance

*Gus Lathouwers, Wieke Harmsen, Catia Cucchiarini, Helmer Strik*
Radboud University

Syllabification concerns the task of dividing words into syllables. Due to many exceptions and subword pattern interactions, training an algorithm to perform syllabification with high accuracy remains a challenge. Different syllabification algorithms have been put forth over the past few decades in the literature, both language-specific and language-independent. Syllabification algorithms can be applied to orthographic representations of words, as well as phonetic representations, and may aid in applications such as text-to-speech and spelling correction software. Given that research on Dutch syllabification algorithms is generally outdated or algorithms are not tailored to Dutch-specific language features, our research set out to apply modern deep-learning techniques for improved syllabification performance. Previously, syllabification algorithms have been applied to phonetic wordsets (e.g., Krantz et al., 2019), and orthographic wordsets (e.g., Trogkanis & Elkan, 2010); yet the two approaches have not been combined to complement each other.

A new deep-learning model was developed that combines orthographic and phonetic information from two independently trained neural nets into a unified deep-learning model using attention mechanisms. Results show that the integration of phonetic in addition to orthographic information in the deep learning model yields improvements. The mean word accuracy of 99.65% is a 0.10% improvement in comparison with the model trained solely on orthographic data, and a 0.14% improvement in comparison with the best model reported in the literature for Dutch orthographic syllabification (Trogkanis & Elkan, 2010). A similar approach using a transformer model applied to the English language achieved a 97.49% word accuracy, representing a 1.18% improvement over the orthographic-only model.

The outcome of the current research indicates that combining phonetic and orthographic information leads to increased accuracy on word processing tasks such as syllabification.

**References**

Trogkanis, N., & Elkan, C. (2010). Conditional random fields for word hyphenation. In J. Hajič, S. Carberry, S. Clark, & J. Nivre (Eds.), Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (pp. 366–374). Association for Computational Linguistics.

Krantz, J., Dulin, M., & De Palma, P. (2019). Language-Agnostic Syllabification with Neural Sequence Labeling. International Conference on Machine Learning and Applications.

# Tonal adaptation of Japanese phonemic loanwords in Mandarin Chinese

*Jueyu Hou (j.hou@hum.leidenuniv.nl)*
Leiden University

The present study accounts for the tonal adaptation of Japanese phonemic loanwords in Mandarin Chinese. Crucially, instead of being dictionary-based, the present study investigated experimentally how 28 selected Japanese loanwords were produced by 12 native Mandarin Chinese speakers (aged 18–35, 5F), none having any (systematic) Japanese L2 learning experience prior to the study. The participants were randomly divided into two groups, each of which was assigned 14 read-aloud tasks (i.e., covering half of the loanwords under study). They were instructed to read aloud the loanwords situated in contextualized declarative carrier sentences. In order to categorize and identify the tonal adaptation strategies, a clustering analysis with *Contour Clustering GUI* was adopted, triangulated with perceptual judgements from two native speakers of Mandarin Chinese. F0 trajectories of the loanwords were extracted with *ProsodyPro* (Xu, 2013) measuring 10 equidistant points per Chinese character to visualize their tones.

The study observed four major tonal adaptation patterns. First, speakers scarcely assign tones to the loanwords following those of the Chinese characters comprising these words. Second, the high-level tone is most commonly assigned, sometimes regardless of the original pitches in Japanese. Third, the assignment of some tones, for instance, the neutral tone, does not always follow the phonological rules of Mandarin Chinese. Finally, between-speaker variations were observed. These findings suggest that despite their readily developed graphic shapes in Mandarin Chinese characters, and despite the fact that the participants have no access to the Japanese lexicon, they tried to “foreignize” the loanwords, pronouncing them in the “Japanese” way as what they imagined to be. These findings made it possible to further discuss the intertwining cognitive and sociolinguistic mechanisms in the process of lexical borrowing.

## Reference

Xu, Y. (2013). ProsodyPro—A tool for large-scale systematic prosody analysis. In B. Bigi & D. Hirst (Eds.), *Proceedings of Tools and Resources for the Analysis of Speech Prosody* (pp. 7–10). Laboratoire Parole et Langage.

# On the possibility of using pre-trained ASR-models to assess oral reading exams automatically

*Bram Groenhof, Wieke Harmsen, Helmer Strik*
Radboud University

Dutch children's reading skills are in decline. One way oral reading skills are measured among primary school students in the Netherlands is the three-minute-exam ('Drie Minuten Toets', DMT). The DMT is time-consuming because teachers must administer it one-on-one, marking word reading correctness in real time. One possible way of alleviating this workload is to use automatic speech recognition (ASR) to aid in the assessment process. Many ASR models struggle with children's speech, but since the DMT only needs a binary correct/incorrect judgment for every word, perfect transcription isn't necessary. Additionally, the ASR-transcriptions can be analysed to obtain diagnostic information about a child's oral reading performance. This information can be utilized by teachers to instruct students based on what type of words or sounds individual students struggle with.

We explored the performance of two state-of-the-art (SOTA) pre-trained ASR-models: Wav2vec2.0 and Whisper. We had them carry out assessments on oral reading word tasks, similar to the DMT, using data from the Children's Oral Reading Corpus (CHOREC). Word lists were read by native Dutch-speaking primary school children aged 6-12 from Flanders and marked by assessors. The judgements of ASR-models and assessors were compared using accuracy, F1-score, and Matthews Correlation Coefficient (MCC) as agreement metrics. Two methods to improve the baseline results were applied: rule-based and similarity-based (using standardized Levenshtein distances).

We found that rule-based improvements obtained the best results for the overall metrics. It involves aspects such as (de)voicing and short versus long vowels. Whisper (accuracy = .54; F1-score = .58; MCC = .54) outperformed wav2vec2.0 (accuracy = .69; F1-score = .39; MCC = .37). The MCC values show that both ASR-models showed mild correlations with assessors.

We conclude that the performance of pre-trained ASR-models, especially Whisper, are promising. We are currently expanding this line of research using recordings of real DMTs. Using the rule-based improvement method, we aim to obtain more detailed diagnostic information from the DMT (e.g., which phonetic aspects the children struggle with).

# *"That's Fantastic!"*
# The Prosodic Characteristics of Sarcasm in Childish Gambino's Speech, Rap, and Singing

*Valerie Querner & Steven Gilbers*
CLCG, University of Groningen

Conversation allows people to establish a sense of connection, and, though seemingly negative, sarcasm plays an important role in this regard (Gibbs, 2000; Recchia et al., 2010). Due to its emotive nature, music is another means of establishing connection (Nummenmaa et al., 2021), and like speech, also features sarcasm (e.g., Bamgbose, 2019). Given that music has been shown to be intricately connected to speech (e.g., Gilbers et al., 2020; Patel et al., 2006), the question arises whether the sarcastic tone of voice functions similarly in music.

The present study explores the prosodic characteristics of sarcasm in speech and music. To this end, a case study was conducted on the speech, rap, and singing of Childish Gambino, comparing sarcastic and non-sarcastic utterances in each domain regarding pitch and rhythm. Based on prior research, it was expected that, in speech, sarcasm would be conveyed through a lower average pitch, less pitch variation, and a slower tempo (e.g., Cheang & Pell, 2008), and that a similar pattern would be observed in music.

The results proved to be mixed: while some similarities remain, musical sarcasm did not mimic the prosody of spoken sarcasm across the board. While sarcastic speech prosody did not differ from sincere speech, sarcastic rap displayed one prosodic cue in line with previous research: a slower tempo. Sarcastic singing is expressed in a different manner altogether. A possible explanation could be the underlying musical structure (e.g., melodic and rhythmic motifs) of the song as a whole. Pitch and rhythm are not the only ways of expressing sarcasm, so artists may opt for other means (e.g., mood juxtaposition of lyrics and instrumentation).

*Keywords: Sarcasm, prosody, language-music connection, singing, rapping.*

**References**


Bamgbose, G. (2020). Beyond rhythm and lyrics: pragmatic strategies of verbal humour in Nigerian hip-hop. *The European Journal Of Humour Research*, *7*(4), 16–31. https://doi.org/10.7592/ejhr2019.7.4.bamgbose

Cheang, H. S., & Pell, M. D. (2008). The sound of sarcasm. *Speech Communication, 50*(5). 366–381. https://doi.org/10.1016/j.specom.2007.11.003.

Gibbs, R. W. (2000). Irony in talk among friends. *Metaphor and Symbol, 15*(1-2), 5–27. https://doi.org/10.1080/10926488.2000.9678862

Gilbers, S., Hoeksema, N., de Bot, K., & Lowie, W. (2020). Regional variation in West and East Coast African-American English prosody and rap flows. *Language and Speech, 63*(4), 713–745. https://doi.org/10.1177/0023830919881479

Nummenmaa, L., Putiken, V., & Sams, M. (2021). Social pleasures of music. *Current opinion in behavioral sciences, 39*, 196–202. https://doi.org/10.1016/j.cobeha.2021.03.026.

Patel, A. D., Iversen, J. R., & Rosenberg, J. C. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French*. Journal of the Acoustical Society of America, 119*(5), 3034–3047. https://doi.org/10.1121/1.2179657

Recchia, H. E., Howe, N., Ross, H. S., & Alexander, S. (2010). Children's understanding and production of verbal irony in family conversations. *British Journal of Developmental Psychology, 28*(2), 255–274. https://doi.org/10.1348/026151008X401903

# Turn-taking in online interactions between people who do and do not stutter

*Lotte Eijk, Stefany Stankova & Sophie Meekings*
Department of Psychology, University of York, Heslington, United Kingdom

Speech seems to effortlessly flow in conversation, with interlocutors timing their utterances based on predictions about the other's speech ([1, 2]). While turn-taking has been well-studied in typical speakers (e.g., [3, 4, 5]), less attention has been given to populations with atypical speech such as people who stutter (PWS). PWS often experience involuntary syllable repetitions, prolongations, and so-called 'blocks' during which speakers are unable to produce sounds. These disfluencies could make their speech less predictable, potentially influencing turn-taking [6]. This study explores turn-taking in conversations with PWS in more detail, focussing on possible differences in turn-taking speed, speaking time, and the likelihood of being interrupted compared to typical speakers.

Twenty conversations were analysed: ten between typical speakers ($M_{age}$ = 29.7), and ten between typical-PWS pairs ($M_{age}$ = 32.8). PWS were self-identified. Speakers participated in a Diapix spot-the-differences task [7] over Zoom. For each of the two rounds, one participant was the leader starting the description, and the other participant the follower.

Preliminary results showed that the leader's turns were longer and role also influenced the type of overlap (automatically coded). We found no evidence for a difference in turn-taking speed, nor for a difference between turn duration or type of overlap between the different speaker groups. These results indicate that negative experiences by PWS could possibly be overcome by giving people clear roles in interactions. Future research could explore this further by using manual coding and investigating the relationship between stuttering severity and turn-taking.

## References


[1] Roelofs, A., & Ferreira, V. S. (2019). The architecture of speaking. *Human language: From genes and brains to behavior*, 35-50.

[2] Meyer, A. S. (2023). Timing in conversation. *Journal of Cognition, 6*(1).

[3] Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics, 38*(4), 555-568.

[4] Templeton, E. M., Chang, L. J., Reynolds, E. A., Cone LeBeaumont, M. D., & Wheatley, T. (2022). Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences, 119*(4), e2116915119.

[5] Bögels, S., Kendrick, K. H., & Levinson, S. C. (2015). Never say no… How the brain interprets the pregnant pause in conversation. *PloS one, 10*(12), e014547.

[6] Freud, D., Moria, L., Ezrati-Vinacour, R., & Amir, O. (2016). Turn-taking behaviors during interaction with adults-who-stutter. *Journal of Developmental and Physical Disabilities, 28*, 509-522.

[7] Baker, R., & Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior research methods, 43*, 761-770.

# A crosslinguistic study on prosodic characteristics in non or minimally verbal autism

*Laura Smorenburg[1], Jill Thorson[2], Aoju Chen[1] and Wolfram Hinzen[3]*

*[1]Utrecht University, [2]University of New Hampshire, [3]Universitat Pompeu Fabra*

Of those diagnosed with autism, approximately one-third is non or minimally verbal (NMVA). Severe impairments are reported in this population for syntax, morphology, and vocabulary in both production (Chenausky et al., 2019) and comprehension (Slušná et al., 2021). One unexplored domain is prosody. Crucially, prosody can be independent of words. Although those with NMVA lack phrase speech and a functional vocabulary, they still vocalize. In the current study, we use existing datasets to ask how prosody is used in the vocalisations of children and adolescents with NMVA from different language environments (American English: Thorson et al., 2016 and Catalan/Spanish: Slušná et al., 2021). American English has a wider pitch range than Catalan and Spanish (Astruc-Aguilera et al., 2009), as well as different syllable structures and rhythmic properties. It was hypothesized that effects of ambient language would be found, implying acquisition of prosodic abilities.

Datasets (English: $N$ = 5, mean age = 12;3, $SD$ = 4;9, Catalan/Spanish: $N$ = 23, mean age = 11;4, $SD$ = 4;0) consisted of the Autism Diagnostic Observation Schedule sessions. All nonvegetative vocalizations were segmented and labelled as 'verbal' or 'nonverbal'. F0 median and span and utterance duration were extracted from each vocalization using Praat (Boersma & Weenink, 2024). Linear mixed modelling was used to look at the effect of ambient language, vocalization type (verbal, nonverbal) and autism severity on prosodic features.

Preliminary results show that verbal vocalizations have lower median F0 than nonverbal vocalizations and that this difference is larger for Catalan/Spanish. Vocalization duration was longer as autism severity increased in English but not Catalan/Spanish, which may be related to the consonant-to-vowel ratio differences between the languages. More analysis is needed to confirm language-specificity in prosody in NMVA. Future work will focus on the communicative function of prosody in NMVA.

**References**


Astruc-Aguilera, L., Payne, E., Post, B., Prieto, P., & Vanrell, M. M. (2009). Acquisition of tonal targets in Catalan, Spanish and English. *Cambridge Occasional Papers in Linguistics*, *5*, 1-14.

Boersma, Paul & Weenink, David (2024). Praat: doing phonetics by computer [Computer program]. Version 6.4.21, retrieved 24 September 2024 from http://www.praat.org/

Chenausky, K., Brignell, A., Morgan, A., & Tager-Flusberg, H. (2019). Motor speech impairment predicts expressive language in minimally verbal but not low verbal, individuals with autism spectrum disorder. *Autism & Developmental Language Impairments, 4,* 1-12. DiStefano, C., & Kasari, C. (2016). The window to language is still open: Distinguishing between preverbal and minimally verbal children with ASD. *Perspectives of the ASHA Special Interest Groups SIG 1*, Vol.1(Part 1).

Slušná, D., Rodríguez, A., Salvadó, B., Vicente, A., & Hinzen, W. (2021). Relations between language, non-verbal cognition, and conceptualization in non-or minimally verbal individuals with ASD across the lifespan. *Autism & Developmental Language Impairments*, *6*. https://doi.org/10.1177/23969415211053264

Thorson, J. C., Usher, N., Patel, R., & Tager-Flusberg, H. (2016). Acoustic analysis of prosody in spontaneous productions of minimally verbal children and adolescents with autism. In the supplement to the proceedings of the *40th Annual Boston University Conference on Language Development (BUCLD)*, Boston, MA.