NVFW
Nederlandse Vereniging voor Fonetische Wetenschappen

Dag van de Fonetiek

ABSTRACTS

6 november 2026

Amsterdam – Kloveniersburgwal 48

Workshopzaal begane grond

Dag van de Fonetiek 2026

vrijdag, 6 november 2026

Kloveniersburgwal 48, Amsterdam

ENTREE = GRATIS

PROGRAMMA

09:30 - 10:00	**Inloop**
10:00 - 10:20	Reduction of Historical Disyllables in Central Kampong Chhnang Cham *Patrick Sam-Lazarov*
10:20 - 10:40	The Prosodic Organization and Implementation of Weak Function Words in Australian English-speaking Children with Speech Sound Disorder *Rui Cai, Paul Boersma, Ivan Yuen, Katherine Demuth, Titia Benders*
10:40 – 11:00	A fully Bayesian feature-based phone recognizer *Orhun Uluşahin, James McQueen*
11:00 - 12:00	**Postersessie 1 (met koffie en thee)**
12:00 - 12:45	**Keynote: Popularization pays: From outreach at festivals via a mobile laboratory to a multimillion euro regional language grant** *Martijn Wieling*
12:45 - 13:00	**Algemene Leden Vergadering**
13:00 - 14:15	**Lunchpauze**
14:15 - 14:35	Do Belgians reverse cross-linguistic trends to maintain phonological contrasts? A study of Belgian French final devoicing – *Jérémy Genette, Jalal Al-Tamimi*
14:35 - 14:55	A quantitative analysis of human corrections to forced alignment and their practical consequences – *Cesko Voeten, Nil Akıntürk*
14:55 - 15:15	Identifying glides in Haitian Creole's determiner allomorphs *Tristan Chopinez*
15:15 - 16:15	**Postersessie 2 (met koffie en thee)**
16:15 - 16:35	Merger or contrast: the distinction between /ɔ/ and /o/ in Paris and Marseille *Eva van Kampen, Heather Burnett, Remco Knooihuizen, Alexander Martin*
16:35 - 16:55	Speech Sound Errors: Categorical or Gradient? *Hugo Quené*
17:00 - 18:00	**Borrel**

Z.o.z. voor postertitels en auteurs

NVFW
Nederlandse Vereniging voor Fonetische Wetenschappen

Posterpresentaties (Doorlopend, tussen 11:00 – 12:00 en 15:15 – 16:15)
Investigation of Norwegian tonal accents and intonation among Norwegian bilinguals *Shayan Aqdas*
The acoustic characteristics of voiced bilabials in Basaá: A three-way contrast of implosives, explosives and nasal elements – *Wiliam Bordonaro, Titia Benders, Silke Hamann, Rodolphe Maah*
Vowel reduction based on prosodic position in Dutch *Tristan Bouverot*
A semi-spontaneous Dutch speech dataset for speech enhancement and speech recognition *Dimme de Groot, Yuanyuang Zhang, Jorge Martinez, Odette Shcaernborg*
Talking to a brick wall? Examining the role of a priori assumptions in computer-directed speech *Jurjen de Haan, Khiet Truong*
The Role of Social Network Size and Diversity in Vowel Identification in Noise *Sarah Kapp, Adriana Hanulíková*
Hearing smiles in the crowd: how babble noise shapes smiled speech perception *Rong Li, Esther Janse, Dirk Heylen, Khiet Truong*
Why can tigers be pigers but babies not dabies? An eye-tracking study investigating the emergence of coronal underspecification in the developing lexicon of 24-months-old Dutch toddlers *Giusy Santo, Merel van Goch, Paula Fikkert*
LLMs at the Phonetics-Phonology-Orthography Interface *Zoe Tholen*
Grammar or Grammars? A maximum entropy analysis of loanword adaptation in Shanghainese *Meng Zhang*

ABSTRACTS

Presentaties

Reduction of Historical Disyllables in Central Kampong Chhnang Cham

Patrick Sam-Lazarov
Recent graduate of the University of Amsterdam

Western Cham is an Austronesian language spoken across Cambodia and southern Vietnam (Eberhard et al., 2026). Tendency towards monosyllables is an areal characteristic of mainland Southeast Asia (Enfield, 2018), and previous literature on Western Cham has documented the weakening of the initial syllable in disyllabic words, resulting in it being dropped or retaining only the onset and forming a word-initial consonant cluster (Friberg & Hor, 1977; Ueki, 2011).

This presentation explores phonetic reduction among historical disyllables in a dialect of central Kampong Chhnang province (CKC), drawing on data from four speakers (all women). In focus are the following three strategies that have not been reported for other Western Cham varieties. For comparison, Proto-Chamic reconstructions by Thurgood (1999) are provided with asterisks:

a. Deletion of the initial syllable with devoicing of the following sonorant. This occurs when the first onset is a stop and the second is a sonorant. Examples include:

1.	a)	[l̥a̤ːw]	*dihlɔw	‘before’
	b)	[m̥ɜ̤ɨ̤]	*tama	‘be cut (by a knife)’
	c)	[ɲ̊iːʔ]	*kuñit	‘turmeric’

b. Deletion of the initial syllable with gemination of the following obstruent. This occurs when both onsets are obstruents. For the coronal fricative /s/, the contrast is reinforced by lenition of the singleton form to a dental approximant [ð̤]. The evidence, though limited, also indicates a length contrast among CKC’s implosives. Examples include:

2.	a)	[pːɛ̤a̤w]	*kabaw	‘water buffalo’
	b)	[kːe̤ː]	*pagi	‘tomorrow’
	c)	[sːaw]	*tasɔw	‘breast’

c. Reduction of an initial syllable with a sonorant into a nasal. This occurs when the first onset is *r, *l or a nasal. For *r and *l, the resulting nasal is homorganic to the following consonant, while for word-initial *m, it either remains [m] or reduces further to a homorganic nasal. In items where the following consonant is already nasal, this reduction produces a geminate. Examples include:

3.	a)	[mpṳʔ]	*mabuʔ	‘drunk’
	b)	[ŋkaj]	*lakɛy	‘man’
	c)	[nːa̤m]	*ranam	‘love (v.)’

Notably, these reduced forms are in synchronic variation with more conservative forms, and there are significant differences among speakers in terms of how often they are produced. Comments from one informant suggest that these differences coincide with diglossia between a ‘spoken’ and ‘written’ form of CKC.

References

Eberhard, D. M., Simons, G. F., & Robinson, A. J (Eds.). (2026). *Ethnologue: Languages of the world*. Twenty-ninth edition. SIL Global. Online version: https://www.ethnologue.com/language/cja/

Enfield, N. J. (2018). *Mainland Southeast Asian languages: A concise typological introduction*. Cambridge University Press. https://doi.org/10.1017/9781139019552

Friberg, T., & Hor, K. (1977). Register in Western Cham phonology. In D. Thomas, E. W. Lee, & D. L. Nguyen (Eds.), *Papers in South East Asian linguistics No. 4: Chamic Studies*. Pacific Linguistics Series A 48, pp. 17–38. https://doi.org/10.15144/PL-A48.17

Thurgood, G. (1999). From ancient Cham to modern dialects: Two thousand years of language contact and change: With an appendix of Chamic reconstructions and loanwords. *Oceanic Linguistics Special Publications* 28, pp. i–407.

Ueki, K. (2011). *Prosody and intonation of Western Cham*. University of Hawai'i at Mānoa dissertation.

The Prosodic Organization and Implementation of Weak Function Words in Australian English-speaking Children with Speech Sound Disorder

Rui Cai[1], Paul Boersma[1],Ivan Yuen[2], Katherine Demuth[3], Titia Benders[1]

[1]University of Amsterdam, [2] Universität des Saarlandes, [3] Macquarie University

Background: Prosodic organization and its phonetic implementation remain relatively understudied in children with speech sound disorder (SSD) (Bernthal et al., 2022), despite their fundamental role in structuring connected speech. Weak function words provide a useful test case because their realization depends on integration into surrounding prosodic structure. Earlier work on typically developing (TD) English-speaking children has identified two relevant patterns of article prosodification: Following a monosyllabic verb, such as *SEE the CAT*, the article may cliticize leftward onto the preceding stressed monosyllabic verb, forming a strong–weak foot. When following a disyllabic verb, as in *FOLlow the CAT*, the preceding word already forms a complete foot, leaving the article unfooted (e.g., Gerken, 1996; Demuth & McCullough, 2009; Redford, 2018). Our previous acoustic studies showed how these patterns are implemented in TD children. Leftward cliticization was reflected in shortening of the preceding verb when an article followed it (Cai et al., 2024), while footed articles are less prominent than unfooted articles (Cai et al., in preparation). It remains unknown whether children with SSD show comparable patterns of prosodic organization and acoustic implementation. The present study therefore asks whether children with SSD show acoustic evidence of leftward cliticization and whether they differentiate between footed and unfooted articles.

Methods: Participants were 19 Australian English-speaking children with SSD (age range: 4;04–12;10; mean age = 7;05) drawn from the AusKidTalk corpus (Ahmed et al., 2021). Two subsets of that project's sentence-repetition task were analysed. The verb-shortening task compared sentences in which a monosyllabic verb was followed by the article *the* (e.g., *Boys often cut the cards*) or followed directly by a noun (e.g., *Boys often cut cards*). The vowel duration in the verb was measured to test whether the presence of the article induced shortening of the preceding verb, providing acoustic evidence of leftward cliticization. The article-prominence task compared the realization of *the* when it followed a monosyllabic verb (e.g., *see the cat*), or a disyllabic verb (e.g., *follow the cat*). Article duration and intensity were combined using principal component analysis to derive a composite measure of article prominence. Linear mixed-effects models tested effects of prosodic condition while accounting for age, speech rate, and task-specific predictors (target verb identity in the verb-shortening task and utterance length for the article-prominence task). The final analyses included 91 productions for the verb-shortening task and 146 for the article-prominence task.

Results: Overall, the findings show partial alignment with patterns previously observed in TD children. In the verb-shortening task, verb durations are shorter when the verb was followed by an article than when followed directly by a noun. In the article-prominence task, however, no reliable difference in article prominence is found between footed and unfooted contexts.

Conclusion: The verb-shortening task finding provides acoustic evidence that children with SSD, like their TD counterparts, cliticize articles leftward and implement this prosodic organization through fine-grained adjustment of the preceding content word. By contrast, children with SSD do not show a reliable footed–unfooted prominence contrast. Two explanations are possible. First, they may differentiate the two contexts, but the present data are insufficient to establish the pattern reliably. Second, the relatively subtle acoustic contrast carried by weak articles may be implemented less consistently in children with SSD. Overall, the findings suggest that children with SSD can establish prosodic organization across word boundaries, while their phonetic implementation do not show the same systematic pattern observed in TD children. Despite the limited sample, these findings broaden our understanding of prosody production in children with SSD and motivate further research on individual variability, SSD subtypes, and the potential value of acoustic measures in clinical assessment.

References

Ahmed, B., Ballard, K.J., Burnham, D., Sirojan, T., Mehmood, H., Estival, D., Baker, E., Cox, F., Arciuli, J., Benders, T. and Demuth, K. (2021). AusKidTalk: an auditory-visual corpus of 3-to 12-year-old Australian children's speech. In *Proc. Interspeech 2021* (pp. 3680-3684).

Bernthal, J. E., Bankson, N. W., Flipsen, P., Fabiano-Smith, L., Gillon, G. T., Goldstein, B. A., Iglesias, A., Justice, L., Kent, R. D., & McLeod, S. (2022). *Speech sound disorders in children : articulation & phonological disorders* (Ninth edition.). Brookes Publishing.

Cai, R., Boersma, P., Yuen, I., Demuth, K., & Benders, T. (2024). Prosodic clitics in English-speaking children's speech production: An acoustic study. In Y. Chen, A. Chen, & A. Arvaniti (Eds.), *Speech Prosody 2024* (pp. 56–60). International Speech Communication Association.

Cai, R., Boersma, P., Yuen, I., Demuth, K., & Benders, T. (2026). *The development of acoustic realization of English articles* [Manuscript in preparation].

Demuth, K., & McCullough, E. (2009). The prosodic (re)organization of children's early English articles. *Journal of Child Language*, *36*(1), 173–200.

Gerken, L. (1996). Prosodic structure in young children's language production. *Language*, *72*(4), 683–712.

Redford, M. A. (2018). Grammatical word production across metrical contexts in school-aged children's and adults' speech. *Journal of Speech, Language, and Hearing Research*, *61*(6), 1339–1354.

A fully Bayesian feature-based phone recognizer

Orhun Uluşahin[1], James M. McQueen[1,2]
[1] Donders Centre for Cognition, [2] Max Planck Institute for Psycholinguistics

Current models of human spoken word recognition (e.g., the Distributed Cohort Model, Gaskell & Marslen-Wilson, 1997, TRACE, McClelland & Elman, 1986, Earshot, Magnuson et al., 2020, and Shortlist B, Norris & McQueen, 2008) rely on a variety of computations to simulate the cognitive mechanisms that underlie human speech processing. However, they share a number of deficiencies across architectures, programming languages, and theoretical frameworks. One such deficiency is ambivalence to speech variation, especially talker-bound variation, which is extensive in real speech. These models all treat speech as if it was spoken by the same talker. A second deficiency is that these models lack the ability to adapt to different forms of speech, including adaptation to talker-specific idiosyncrasies (Eisner & McQueen, 2005).

We present a new model of spoken word recognition, the Adaptive Bayesian Continuous-speech (ABC) Model that does not have these two deficiencies. The model builds on the Bayesian framework of Shortlist B's word recognition process but extends its probabilistic mode of operation to a feature-based Bayesian phone recognizer as well. Thus, instead of relying on hard-coded confusion data from human participants (i.e., as in Shortlist B), ABC dynamically evaluates pseudo-spectral speech input (i.e., numerical representations of acoustic features) of each incoming segment on a feature-by-feature basis to compute relative probabilities for each phone in the listener's inventory. In ABC, phones are a combination of position-specific allophones (where necessary to model positional variability) and position-ambivalent phonemes, in line with allophonic processing at the pre-lexical level (Mitterer et al., 2013, 2018). Knowledge about different talkers' acoustic-phonetic idiosyncrasies is represented in the form of Gaussian or Beta distributions that constitute phone-level priors on a feature-by-feature basis. Thus, ABC enables the simulation of talker-specific perception, accommodating the "talker-contingent" nature of speech recognition (Nygaard et al., 1994). In doing so, our model expands the limited scope of feature-based Bayesian phone perception (e.g., Kleinschmidt & Jaeger, 2015; Xie et al., 2023) in two substantial ways: (1) ABC uses a language-sized inventory of allophones instead of an individual phoneme pair, and (2) compares an expandable array of 12 acoustic features instead of relying on one acoustic-phonetic contrast.

The combination of the Bayesian framework and a modular architecture keep computational demand to a minimum while maximizing user customizability. For instance, users can simulate different "listening profiles" based on between- or within-talker differences. Perceptual expectations can be configured simply by altering the feature-by-feature distributions that constitute phone priors. The lexicon can be expanded, shrunk, or completely replaced as long as it follows ABC's barebones format. Acoustic features can be added or removed so long as they can be represented by a Gaussian or Beta distribution.

Therefore, although the primary use case for the ABC phone recognizer is its integration within the larger model of word recognition, the phone recognizer module alone has a number of other use cases. These include (1) Exploring probabilistic/graded phone perception as opposed to the strict categorical perception of many models, (2) Modelling cue weighting through changes to feature mismatch penalties and feature-by-feature weights, and (3) Simulation of degraded, clinical or otherwise atypical speech or listening conditions.

At Dag van de Fonetiek 2026, we will present a detailed overview of the probabilistic phone recognizer module of ABC, and present simulation results that capture (talker-bound) perceptual learning (Eisner & McQueen, 2005) and its generalization across the lexicon (McQueen et al., 2006) with a phonetic inventory of 49 Dutch phones in a lexicon of 35,000 words. These simulations will demonstrate ABC's adaptability, expandability, and its capacity to handle acoustic variation in speech, with a particular emphasis on between-talker variability, addressing multiple major deficiencies in existing models of spoken word recognition.

References

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238. https://doi.org/10.3758/BF03206487

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating Form and Meaning: A Distributed Model of Speech Perception. *Language and Cognitive Processes*, *12*(5–6), 613–656. https://doi.org/10.1080/016909697386646

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203. https://doi.org/10.1037/a0038695

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., & Rueckl, J. G. (2020). EARSHOT: A Minimal Neural Network Model of Incremental Human Speech Recognition. *Cognitive Science*, *44*(4), e12823. https://doi.org/10.1111/cogs.12823

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological Abstraction in the Mental Lexicon. *Cognitive Science*, *30*(6), 1113–1126. https://doi.org/10.1207/s15516709cog0000_79

Mitterer, H., Reinisch, E., & McQueen, J. M. (2018). Allophones, not phonemes in spoken-word recognition. *Journal of Memory and Language*, *98*, 77–92. https://doi.org/10.1016/j.jml.2017.09.005

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition*, *129*(2), 356–361. https://doi.org/10.1016/j.cognition.2013.07.011

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech Perception as a Talker-Contingent Process. *Psychological Science*, *5*(1), 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Xie, X., Jaeger, T. F., & Kurumada, C. (2023). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review. *Cortex*, *166*, 377–424. https://doi.org/10.1016/j.cortex.2023.05.003

Funding

This research is supported by the Dutch Research Council (NWO), Social Sciences and Humanities Open Competition grant 406.23.CTW.005.

KEYNOTE

Popularization pays: from outreach at festivals via a mobile laboratory to a multimillion euro regional language grant

Martijn Wieling
Rijksuniversiteit Groningen

I like going to (music) festivals. This was probably not the sentence you were expecting when reading an abstract of an invited speaker for a conference. However, in my presentation I will explain why this sentence in fact is quite important for the work in our lab investigating pronunciation variation. Specifically, I will try to argue why festivals do not only make a good environment for popularizing research, but also make an excellent environment for collecting speech production data in a broad population. In particular, I will highlight a study (Wieling et al., 2020) conducted at the festival Lowlands in 2019 during which we investigated accomodation in speech production, a potential driver of language change (Trudgill, 2008). I will also touch upon an acoustic-only technique we developed to quantify pronunciation differences which performs on par with human judgements (Bartelds et al., 2022; Bartelds & Wieling, 2022).

To facilitate our speech data collection efforts and to simultaneously increase our visibility when collecting data, we were very lucky in that my university has facilitated the acquisition of a one-of-a-kind mobile laboratory containing a sound-dampened room (Rebernik et al., 2025). Since this mobile laboratory, dubbed SPRAAKLAB, has also made collecting research-grade speech production data in noisy environments feasible, we have increased our data collection efforts at various festivals in the Netherlands and abroad. In my presentation, I will discuss the characteristics of this unique research tool.

I will then discuss how SPRAAKLAB has led to a substantial (in principle 30-year) regional language revitalization grant, resulting in the project "MOI: Meertalig Onderwijs In Grunnen en Drenthe" (www.wijzijnmoi.nl) in which we aim to integrate Gronings and Drents in primary schools, and support both teachers and pupils through speech and language technology.

Time permitting, I will also touch upon some novel initiatives to interest a general audience in our work, including our MOOC for Gronings, our board game Streektaalstrijd, and a newly developed Escape Classroom Activity about the "Witte Wieven".

Do Belgians reverse cross-linguistic trends to maintain phonological contrasts? A study of Belgian French final devoicing.

Jérémy Genette[1], Jalal Al-Tamimi[2]
[1]KU Leuven, [2]Université Paris Cité, CNRS, LLF

This contribution investigates the phonetic realization of phonological voicing contrasts in the fricatives of Belgian French, compared with Hexagonal French. Maintaining voicing at the end of prosodic units is cross-linguistically challenging (e.g., Blevins, 2006). As French is a prototypical true-voicing language (e.g., Beckman, 2013) and allows both word-final stops and voicing contrasts in word-final position, French provides an ideal testing ground for investigating the mechanisms underlying the maintenance of voicing contrasts in final position (e.g., Jatteau, 2025).

To maintain voicing, speakers can circumvent the *Aerodynamic Voicing Constraint* by resorting to different strategies: (i) lengthening the preceding vowel, (ii), reducing the duration of the fricative or (iii) increasing the V/VC ratio (Ohala, 2011). These strategies have been shown cross-linguistically to support the maintenance of voicing in fricatives, particularly in final position. However, Belgian French has been described as a variety in which vowels are lengthened before fricatives when the fricatives are devoiced, with fricatives tending to be devoiced in syllable-final position (Hambye, 2005). However, no lengthening is expected when the fricative is voiceless. This thus raises the question of how durational and voicing cues are weighted and how they interact to convey the voicing contrast in Belgian French fricatives.

To answer this question, fricatives in syllable-final position were extracted from two corpora: the *Corpus de Référence du Français Parlé* for Hexagonal French (Delic, 2004; 36 hours of speech) and *VALIBEL* for Belgian French (Francard et al., 2009; 43 hours of speech). The data were force-aligned using the Montreal Forced Aligner (McAuliffe et al., 2017). A total of 24,867 fricatives in final position from 540 speakers were analyzed. The duration of the fricative, the duration of the preceding vowel, and the V/VC ratio were extracted. The voicing properties of the fricatives were assessed following the two-pass filtering method described in Al-Tamimi & Khattab (2018), allowing us to calculate the proportion of each fricative exhibiting voicing. All cues were speaker-normalized, and their weights were estimated via logistic regression.

The results show a significant interaction between country (Hexagonal vs Belgian French) and the weights assigned to duration and voicing cues. Marginal-effects analyses reveal that French speakers tend to follow the expected cross-linguistic pattern. The more a fricative is phonetically devoiced: i.e., the lower the proportion of vocal-fold vibration during the fricative, the shorter the preceding vowel, as shown by the increasingly lower orange line across the panels from left to right in Figure 1. In Belgian French, however, the pattern appears to be reversed: when a final fricative is devoiced, a longer preceding vowel is associated with a higher probability that the fricative is phonologically voiced (left panel in Figure 1). Conversely, when the fricative is voiced, the probability that it is phonologically voiceless also increases as the duration of the preceding vowel increases.

In conclusion, the results show that, in Hexagonal French, the correlation between voicing and durational cues is consistent with the phonetic principles used to maintain voicing. In contrast, the patterns observed in final fricatives in Belgian French are not accounted for by the same cue-weighting relationship. Belgian French speakers appear to encode phonological voicing in the duration of the preceding vowel and, consequently, in the V/VC ratio, making greater use of this cue when the fricative is devoiced. This suggests a cue-trading relationship between voicing and vowel duration (see Repp, 1982; Clayards, 2017).

References

Al-Tamimi, J., & Khattab, G. (2018). Acoustic correlates of the voicing contrast in Lebanese Arabic singleton and geminate stops. *Journal of Phonetics, 71*, 306–325. https://doi.org/10.1016/j.wocn.2018.09.010

Beckman, J., Jessen, M., & Ringen, C. (2013). Empirical evidence for laryngeal features: Aspirating vs. true voice languages. *Journal of Linguistics, 49*(2), 259–284. https://doi.org/10.1017/S0022226712000424

Blevins, J. (2006). Theoretical synopsis of Evolutionary Phonology. *Theoretical Linguistics*, 32(2), 117–166.

Clayards, M. (2017). Individual talker and token covariation in the production of multiple cues to stop voicing. *Phonetica*, 75(1), 1–23. https://doi.org/10.1159/000448809

Delic, É. (2004). Présentation du Corpus de référence du français parlé. *Recherches sur le français parlé*, 18, 11–42.

Francard, M., Hambye, P., Simon, A. C. et Dister, A. (2009). Du corpus à la banque de données : Du son, des textes et des métadonnées. L'évolution de banque de données textuelles orales VALIBEL (1989-2009). *Cahiers de l'Institut de linguistique de Louvain-CILL*, 33(2), 113-29.

Hambye, P. (2005). *La prononciation du français contemporain en Belgique: variation, normes et identités* (Doctoral dissertation, UCLouvain).

Jatteau, A., Audibert, N., Vasilescu, I., Lamel, L., & Adda-Decker, M. (2025). Final devoicing before it happens: A large-scale study of word-final obstruents in French. *Laboratory Phonology, 16*(1). https://doi.org/10.16995/labphon.10855

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Interspeech 2017* (pp. 498–502). ISCA.

Ohala, J. (2011). Accommodation to the aerodynamic voicing constraint and its phonological relevance. In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 64–67).

Repp, B. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, *92*(1), 81–110. https://doi.org/10.1037/0033-2909.92.1.81

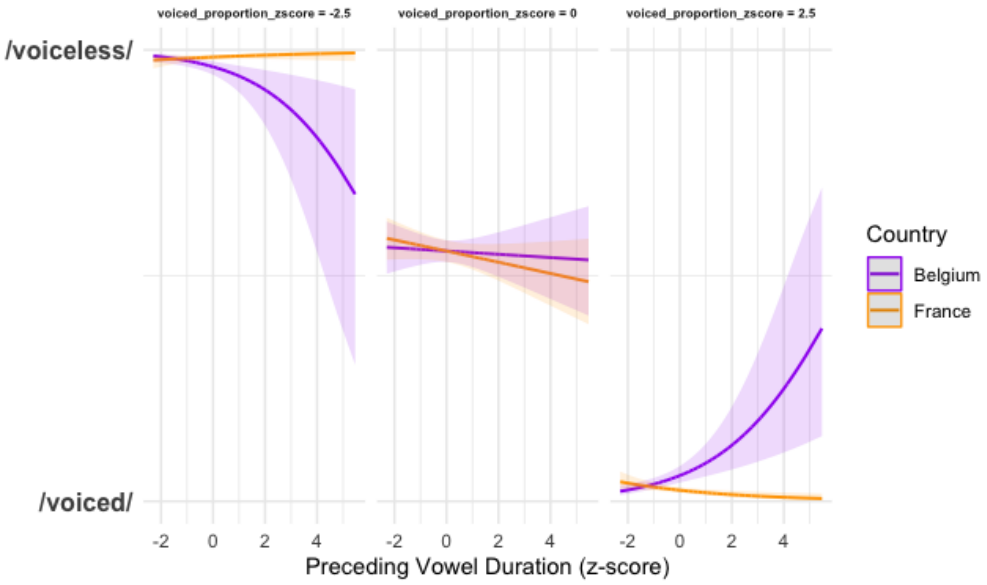


Figure 1: Predicted marginal effects of vowel duration and proportion of phonetic voicing on phonological voicing probability in Belgian and Hexagonal French.

A quantitative analysis of human corrections to forced alignment and their practical consequences

Nil Akintürk[1], Cesko Voeten[12]
[1]University of Amsterdam, Amsterdam Center for Language and Communication, [2]Fryske Akademy

A common approach to segmenting speech into phones is to use *supervised forced alignment*, where recordings are manually annotated at the sentence or word level, and a computer tool takes these annotations (in IPA, usually through G2P conversion) and aligns the phone boundaries to the speech signal. Both conventional wisdom and hard data hold that automated forced alignment should be hand-corrected post-hoc, due to inevitable errors. However, manual corrections are highly laborious; what's more, they are often done by multiple correctors (e.g. different student assistants), introducing additional variation.

The present paper is focused on this second source of variation. While it is well-established that a dataset with hand-corrections is superior to an uncorrected ('baseline') dataset (e.g. MacKenzie & Turton 2020), we extend this by studying variation *among* correctors. In addition, we assess the *practical consequences* for research in the domain of phonetic variation in change. Evanini, Isard, & Liberman (2009) have shown for *automatic formant extraction* that if a corpus is sufficiently large, formant-tracking errors are negligible in practice; we examine whether this might hold similarly for our research object of phone segmentation. Here, too, we explicitly take into account the additional factor of inter-corrector variation.

Our dataset consists of a total of 112,058 vowel tokens, of which 58,714 had been corrected by a total of six student assistants (all Linguistics majors with ample background in phonetics) plus their supervisor (an assistant professor of sociophonetics/sociophonology). These (non-)corrections were obtained from the Boarnsterhim Corpus (Kingma et al submitted), a recently-completed dataset containing sociolinguistic-interview and word-list data from 112 speakers from the Frisian municipality of Boarnsterhim. The corpus had been manually transcribed orthographically and aligned at the sentence level, and subsequently been aligned at the word and, in a second pass, phone levels using the Montréal Forced Aligner (McAuliffe et al 2017; note that this aligner is currently the top state of the art according to the comparison by Rousso et al 2024).

Our analysis consisted of both exploratory analysis and confirmatory hypothesis testing. Exploratorily, we investigated (i) the size of the disagreement between the aligner's automated boundaries and human corrections stratified by manner of articulation, (ii) the extent to which correctors disagree with one another on the same boundary type stratified by manner of articulation, and (iii) whether formant transitions at these boundaries can explain these disagreements. Confirmatory hypotheses tests assessed (iv) the extent to which corrections of the starting and end points of the same tokens significantly differed among correctors (using LMMs); (v) whether specific formant transitions predicted these corrections; (vi) whether speaker characteristics (specifically, sex and speaker identity) and vowel characteristics (position, consonant voicing, consonant PoA, consonant manner, vowel height, vowel backness) predicted these corrections; (vii) whether the corrections have *practical consequences* for the material conclusions that would be drawn by a sociophonetic analysis of these data's F1, F2, and F3 values (using GAMMs [Wood 2017], cf. Voeten, Heeringa, & Van de Velde 2022).

Our results show that both aligner-corrector and inter-corrector disagreement are largest for sonorants and vowels that are bordering silence, and the smallest where the vowel meets an obstruent, suggesting that there is an overlap between the cues that are used by the model and the humans annotators. Significant differences among human correctors were found; they are not significantly associated with formant transitions, but various effects of speaker and vowel characteristics are significantly associated with these inter-corrector differences. Perhaps most importantly, there were very little differences in GAMM fits between uncorrected data, corrected data overall, and corrected data split out by the different correctors. All of this suggests that while human corrections may be superior to unsupervised forced alignment, the high cost of human corrections may not justify the modest gains.

References

Evanini, K., Isard, S., & Liberman, M. Y. (2009). Automatic formant extraction for sociolinguistic analysis of large corpora. In *Proceedings of Interspeech 2009* (pp. 1655-1658).

MacKenzie, L., & Turton, D. (2020). Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard*, *6*(s1), 20180061.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of Interspeech 2017* (pp. 498-502).

Kingma, M., Pinget, A-F., Heeringa, W. & Van de Velde, H. (submitted). The Boarnsterhim Corpus: A Frisian-Dutch bilingual speech corpus in apparent- and real-time. *Language Resources and Evaluation*.

Rousso, R., Cohen, E., Keshet, J., & Chodroff, E. (2024). Tradition or innovation: A comparison of modern ASR methods for forced alignment. *arXiv preprint arXiv:2406.19363*.

Voeten, C. C., Heeringa, W., & Van de Velde, H. (2022). Normalization of nonlinearly time-dynamic vowels. *The Journal of the Acoustical Society of America*, *152*(5), 2692-2710.

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd edn.). Chapman & Hall/CRC.

Identifying glides in Haitian Creole's determiner allomorphs

Tristan Chopinez
Universiteit Leiden

The Haitian Creole determiner morpheme displays an unusual allomorphy, surfacing as [la] following consonant-final nouns, and as [a] following vowel-final nouns (Bonet et al., 2007; Cadely, 2002; Klein, 2003; Lahrouchi & Ulfsbjorninn, 2024; Lefebvre & Fournier, 1979; Nikiema, 1999; Nikiema & Bhatt, 2003; Picard, 2001), as in table 1. This pattern maximizes markedness in all contexts, as opposed to either reducing markedness (markedness constraints) or ignoring markedness (faithfulness constraints). This phenomenon, which we can call anti-markedness, is difficult to account for theoretically, as shown by Klein (2003) in tableaux 1 & 2.

According to the literature (Bonet et al., 2007; Cadely, 2002; Klein, 2003; Nikiema, 1999; Nikiema & Bhatt, 2003; Picard, 2001; Valdman, 1978, 2015), this anti-markedness process is disturbed by a secondary allomorphic process: glide insertion after a (mid-)high noun, as in table 2. If correct, this complicates the morphophonological analysis of this morpheme. Haitian selects maximally marked forms, only to reduce markedness in some of these forms. Whichever mechanism can explain Haitian's anti-markedness must thus be violated within this same environment.

However, this insertion is not described in any standard description or learning materials for Haitian, it is instead only stated to occur in a handful of studies. Most of these studies are citing other works (Bonet et al., 2007; Cadely, 2002; Klein, 2003; Nikiema, 1999; Nikiema & Bhatt, 2003; Picard, 2001). These cited works either themselves cite other works, or merely claim this to be true without thorough justification (Valdman, 1978, 2015). Given the general acoustic similarities between glides and high-vowel transitions, it cannot be ruled out that the claims come from a misinterpretation of what researchers heard when interacting with Haitian speakers. Due to this uncertainty, paired with the theoretical relevance, a detailed investigation of the vowel hiatus regions in Noun-Determiner segments in Haitian must be undertaken.

Spoken data from the Radio Haïti corpus (Havard et al., 2025) is analyzed for this investigation. Contexts with high-vowel-final nouns are compared to ones with mid-high-vowel-final nouns. If there is no glide, it would be expected that the F1 of these vowels would simply rise towards [a]. If there is a glide, there would not be a very pronounced change in the F1 from [i] to [j] and from [u] to [w], before it rising towards [a]. However, there would be a notable drop in F1 from [e] to [j] and from [o] to [w], prior to it rising towards [a].

To test this, 159 individual noun-final vowels to determiner [a] transitions are found in the corpus. In Praat (Boersma & Weenink, 2025), they are first each broken into 100 segments. A measurement of the F1 is then taken at each segment boundary (for 101 total measurements per transition). Two methods are used to analyze the data. First, a general comparison of the F1 trajectories of the two groups is conducted using a general additive model (GAM) in RStudio (Wood, 2017). The GAM uses the measurements to reconstruct each formant trajectory, and compares the different trajectories on the basis of the first vowel's quality (90 have high vowels, 69 have mid-high vowels), to look for any non-random deviance between the two groups (Hastie & Tibshirani, 1987). This reveals significant differences in the trajectories. Second, the amount of the trajectory that is below the starting point (F1 dip) is measured for each F1 trajectory. These measurements are compared between groups using a linear model. No significant difference is found between them, and 70.7% of mid-high vowel sequences feature no F1 dip at all. Thus, although there are statistically meaningful differences in the shape of the trajectory of the F1 in [ia] & [ua] versus [ea] & [oa], these do not correspond to the expected effects an apparent glide between the noun-final vowel and the determiner's [a].

A second study is currently underway, using the same corpus to compare Noun-Determiner environments to word-internal Vowel-Glide-Vowel environments, and using the same GAMs analysis method (with Praat and R) to test for differences in those trajectories. If there are no glides in the determiner environment, then it is expected that the F1 trajectories from this context will be different to those in the word-internal Vowel-Glide-Vowel environment.

Figures

Table 1: Haitian Determiner Allomorphy

/liv + DET/	/papa + DET/
[livla]	[papaa]
the book	*the father*

Adapted from Klein (2003).

Table 2: Haitian Determiner Allomorphy with Hiatus

/lapli + DET/	/papje + DET/	/vɛ + DET/	/tu + DET/	/bato + DET/	/bɔkɔ + DET/
[laplija]	[papjeja]	[vɛa]	[tuwa]	[batowa]	[bɔkɔa]
the rain	*the paper*	*the glass*	*the hole*	*the boat*	*the sorcerer*

Adapted from Klein (2003).

Tableaux 1 & 2

Markedness and Faithfullness constraints leading to the wrong outputs

papa+la	Onset	NoCoda	Max
a. ☞ pa.pa.la			
b. pa.pa.a	∗!		∗

liv+la	Onset	NoCoda	Max
a. liv.la		∗!	
b. ☞ li.va			∗

*A*dapted from Klein (2003).

References

Boersma, P., & Weenink, D. (2025). Praat: Doing phonetics by computer. https://www.praat.org

Bonet, E., Lloret, M.-R., & Mascaró, J. (2007). Allomorph selection and lexical preferences: Two case studies. *Lingua*, *117*, 903–927. https://doi.org/10.1016/j.lingua.2006.04.009

Cadely, J.-R. (2002). Le statut des voyelles nasales en créole haïtien. *Lingua*, *112*, 435–464.

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: Some applications. *Journal of the American Statistical Association*, *82*(398), 371–386.

Havard, W. N., Govain, R., & Benjamin Lecouteux, E. S. (2025). Self-supervised models of speech processing for haitian creole. *Interspeech*, 4018–4022.

Klein, T. B. (2003). Syllable structure and lexical markedness in creole morphophonology: Determiner allomorphy in haitian and elsewhere. In I. Plag (Ed.), *Phonology and morphology of creole languages* (pp. 209–228). Niemeyer.

Lahrouchi, M., & Ulfsbjorninn, S. (2024). Nasal assimilation counterfeeding and allomorphy in haitian: Nothing is still something! *Linguistic Inquiry*, *55*(2), 255–286. https://doi.org/10.1162/ling_a_00469

Lefebvre, C., & Fournier, R. (1979). La particule la en créole haïtien. *Cahier de linguistique*, (9), 37–72. https://doi.org/10.72-2/800077ar

Nikiema, E. (1999). De la variation du déterminant /la/ dans les créoles haïtien et st-lucien. *Lingua*, *107*, 69–93.

Nikiema, E., & Bhatt, P. (2003). Two types of r deletion in haitian creole. In I. Plag (Ed.), *Phonology and morphology of creole languages* (pp. 43–69). Niemeyer.

Picard, M. (2001). Aspects synchroniques et diachroniques de l'hiatus: Le cas du déterminant /la/ en créole haïtien. *Revue québécoise de linguistique*, *30*(2), 99–112. https://doi.org/10.7202/000521ar

Valdman, A. (1978). *Le créole: Structure, statut et origine*. Éditions Klincksieck.

Valdman, A. (2015). *Haitian creole: Structure, variation, status, origin*. Equinox.

Wood, S. N. (2017). *Generalized Additive Models: An introduction with R* (2nd ed.). Chapman; Hall/CRC.

Merger or contrast: the distinction between /ɔ/ and /o/ in Paris and Marseille

Eva van Kampen[1], Heather Burnett[2], Remco Knooihuizen[1], Alexander Martin[1]
[1] University of Groningen, [2] Université Paris Cité

Standard French distinguishes between the mid vowels /ɔ/ and /o/, which creates minimal pairs such as *pomme* 'apple' /pɔm/ vs. *paume* 'palm of the hand' /pom/. Earlier studies have identified a shift in the relation between these two phonemes as their realisations appear to be converging (Cecelewski et al., 2014, Hansen, 2011). However, there is variability between speakers at the regional and individual levels in whether these are distinct categories: Avanzi (2017) reported regional perceptual differences between speakers from the North and the South of France, with predominantly distinct categories in the North and merged categories in the South. Weng and Martin (in prep.), however, show that speakers *within* each region differ in their production of these vowels. In light of this intra-regional variation, they argue for a stronger focus on the level of the individual. Despite this observation, there is still no detailed description of the production of the vowels /ɔ/ and /o/ that takes this variability into account. Such a description at the individual level is, however, necessary to provide insight into the development of changes at the community level (Labov, 2001). Therefore, the objective of the current study is to map the extent of individual-level variation in the production of these vowels as a first step towards further research.

Specifically, this project aims to map the differences between speakers in two regions of France in order to assess vowel variation in major French cities: Paris in the north and Marseille in the south. For this purpose, we use the CaFé corpus (*Cartographie linguistique des féminismes*, Abbou & Burnett, 2024), which contains sociolinguistic interviews with 40 speakers from Paris and 20 speakers from Marseille. We have created an automated analysis pipeline using the tools Pyannote (Plaquet & Bredin, 2023), Whisper (Radford et al., 2022) and the Montreal Forced Aligner (McAuliffe et al., 2017), which processes audio files to identify, extract and analyse target vowels. This method allows us to analyse data from a corpus of natural speech (approximately 90 hours) without manual intervention.

Our analysis focuses on the frequencies of the first and second formants (F1 and F2, following Storme, 2017), which allows us to determine whether a speaker produces a reliable acoustic contrast between the two vowels using the Pillai score (following Stanley & Sneller, 2023). These scores allow us to take the first steps in a socio-phonetic analysis of the variation found in the production of /ɔ/ and /o/ on an individual level. At the Dag van de Fonetiek, we will go into detail on the variability between individual speakers as well as on the geographical dimension.

References

Abbou, J., & Burnett, H. (2024). Devenir féministe à Paris et Montréal: Récits de vie dans le corpus CaFé. In press in H. Blondeau, W. Remysen & M. Laforest (Eds.), *80 ans des corpus montréalais*. Presses universitaires de Montréal.

Avanzi, M. (2017). *Atlas du français de nos régions*. Paris: Armand Colin.

Cecelewski, J., Gendrot, C., Adda-Decker, M., & Boula de Mareüil, P. (2024). Diachronic evolution of /e ɛ o ɔ/ in French across 100 years of speech archives. *Journal of French Language Studies*, *34*(3), 483–515. doi:10.1017/S0959269524000206

Hansen, A. B., & Juillard, C. (2011). La phonologie parisienne à trente ans d'intervalle–Les voyelles à double timbre. *Journal of French Language Studies*, *21*(3), 313-359.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

Plaquet, A., & Bredin, H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. *INTERSPEECH 2023*, 3222–3226. https://doi.org/10.21437/Interspeech.2023-205

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2212.04356

Stanley, J. A., & Sneller, B. (2023). Sample size matters in calculating Pillai scores. *The Journal of the Acoustical Society of America, 153*(1), 54–67. https://doi.org/10.1121/10.0016757

Storme, B. (2017). The loi de position and the acoustics of French mid vowels. *Glossa: a journal of general linguistics*, 2(1), 64.

Weng, C. & Martin, A. (in prep.) Cross-dialect perception and production of the French mid vowels in mobile speakers.

Speech sound errors: Categorical or gradient?

Hugo Quené
Utrecht University

Speech sound errors provide valuable insights into how speakers mentally prepare their speech. Speakers detect and repair some (but not all) errors in their own speech. This detection may happen relatively early, or even immediately (e.g. *v--horizontal*), or it may happen relatively late (*go from left again... eh, from pink again to blue*). The duration of the error-to-interruption interval follows a bimodal distribution, with the peaks separated by about 350 ms (Nooteboom & Quené, 2013). This suggests two different detection processes, and hence two different stages of mental preparation of speech. Presumably, the first (early) detection taps into abstract phonological processing, whereas the second (late) detection taps into articulatory preparation and/or execution. Hence, it is hypothesized that early-detected errors tend to be categorical, that is, without a phonetic trace of the target sound, whereas late-detected and undetected errors tend to be gradient, that is, with articulatory blending resulting in a phonetic trace of the intended target into the produced sound.

This hypothesis was investigated using data from two sources: (1) elicited Dutch sound errors from several experiments involving phonological priming (Quené & Nooteboom, 2024), as well as (2) a sample of spontaneous English sound errors from the Simon Fraser University Speech Error Database (SFUSED, based on English podcasts; Alderete, 2025; Alderete et al., 2021). For both corpora, consonants were taken from correct controls, from early-detected errors, from late-detected errors (Dutch only), and from undetected errors.

The Dutch errors were studied perceptually *via* a consonant identification task using human listeners; the results of this perceptual study confirm the main hypothesis (Nooteboom & Quené, 2013). The English errors were studied acoustically by means of measurements of the Voice Onset Time (VOT) of critical consonants; the results of this acoustic analysis are inconclusive. Although the VOT data are consistent with the main hypothesis, there were too few early-detected errors (n=13) in the sample to corroborate a null effect, and the trace effect in the undetected errors (n=73) was relatively small. An additional sample of 242 English speech errors from the SFUSED also failed to corroborate the predicted pattern in VOT.

In conclusion, the combined results suggest that speech sound errors that are detected early seem to be categorical, without any phonetic trace of the intended speech sound, whereas sound errors detected late seem to be more gradient. The evidence is clear in the Dutch elicited errors, and less convincing for English spontaneous speech errors. These findings suggest that speakers have different mental representations of their speech plan in the earlier vs. later stages of mental preparation of speech.

References

Alderete, J. (2025). Simon Fraser University Speech Error Database – English (SFUSED English): Methods and Design. *Journal of Cognition*, *8*(1), 28. https://doi.org/10.5334/joc.440

Alderete, J., Baese-Berk, M., Leung, K., & Goldrick, M. (2021). Cascading activation in phonological planning and articulation: Evidence from spontaneous speech errors. *Cognition*, *210*, 104577. https://doi.org/10.1016/j.cognition.2020.104577

Nooteboom, S.G. & Quené, H. (2013). Parallels between self-monitoring for speech errors and identification of the misspoken segments. *Journal of Memory and Language*, 69 (3), 417–428. doi:10.1016/j.jml.2013.04.006, Corrigendum (2023): doi:10.1016/j.jml.2023.104448.

Quené, H., & Nooteboom, S.G. (2024). Some properties of mental speech preparation as revealed by self-monitoring. *Speech Communication*, 103043. https://doi.org/10.1016/j.specom.2024.103043

Acknowledgements

My sincere thanks are due to Sieb Nooteboom, for his support and guidance in our past work on speech errors, and for his critical comments on earlier versions of this presentation. In addition, I am very grateful to Matt Goldrick and John Alderete for inspiring discussions about this work, for sharing the full SFUSED corpus of speech errors, and for sharing their VOT data (Alderete et al., 2021).

References

Alderete, J. (2025). Simon Fraser University Speech Error Database – English (SFUSED English): Methods and Design. *Journal of Cognition*, 8(1), 28. https://doi.org/10.5334/joc.440

Alderete, J., Baese-Berk, M., Leung, K., & Goldrick, M. (2021). Cascading activation in phonological planning and articulation: Evidence from spontaneous speech errors. *Cognition*, 210, 104577. https://doi.org/10.1016/j.cognition.2020.104577

Nooteboom, S. G., & Quené, H. (2013). Parallels between self-monitoring for speech errors and identification of the misspoken segments. *Journal of Memory and Language*, 69(3), 417–428. https://doi.org/10.1016/j.jml.2013.04.006

Quené, H., & Nooteboom, S. (2024). Some properties of mental speech preparation as revealed by self-monitoring. *Speech Communication*, 103043. https://doi.org/10.1016/j.specom.2024.103043

ABSTRACTS

Postersessie

(in alfabetische volgorde)

Investigation of Norwegian tonal accents and intonation among Norwegian bilinguals

Shayan Aqdas[1]
[1]University of Stavanger, Norway

The study investigates the acquisition and representation of Norwegian tonal accents, with a focus on bilingual speakers. Norwegian is characterized by a surface contrast between two tonal accents, which are anchored in word stress. Stressed syllables in Norwegian can be realized in two ways, resulting in Accent 1 or Accent 2, and which in some cases can differentiate word meanings through distinct tonal patterns on stressed syllables, such as 1*tanken* 'the tank' vs. 2*tanken 'the thought'*. The phonetic realization of tonal accents may also vary by dialect.

Situated at the intersection of heritage language linguistics, second language acquisition, and phonological theory, the study adopts a three-way comparative design involving L1 Norwegian speakers, heritage bilinguals with Urdu as a heritage language, and late L2 learners of Norwegian with Urdu as their first language. Urdu provides a relevant comparison language, as it has word stress but does not distinguish between tonal accents (Mumtaz, Bögel, & Butt, 2020). The project is theoretically grounded in the Micro-Cue Model (Westergaard, 2019). This framework helps to investigate how grammatical features are acquired or restructured under second language acquisition.

The project combines perception and production experiments. The categorial perception of stress placement and tonal accent distinctions will be assessed using ABX discrimination tasks. Production will be elicited through structured sentence-level tasks based on Hognestad's (2012) structured elicitation of words containing Accent 1 and 2. Acoustic analyses will focus on fundamental frequency contours, tonal alignment, and prosodic timing; these patterns will be assessed in relation to participants' background characteristics.

It is hypothesized that L1 Norwegian speakers will perform near ceiling on all tasks. Heritage Urdu speakers might show high accuracy in perception of tonal accents if early exposure attuned them to this contrast. Late L2 learners, however, are expected to find tonal accent discrimination more challenging, potentially testing at chance. The three groups are expected to successfully distinguish stress placement, as both languages encode word-level stress. Contrastively, tonal accents will pose challenges for the groups, especially, bilingual groups. If learners show reliable perception and production of Accent 1 and 2, it will indicate a stable grammatical representation of tonal accents, however, if they perceive without producing target contrasts or vice versa, it will indicate their reliance on routine use or imitation rather than an established grammatical representation.

References

Hognestad, J. K. (2012). Tonelagsvariasjon i norsk. Universitetet i Agder

Mumtaz, B., Bögel, T., & Butt, M. (2020). Lexical Stress in Urdu. *Interspeech*, 1888-1892. doi:http://dx.doi.org/10.21437/Interspeech.2020-2942

Westergaard, M. (2019). Microvariation in Multilingual Situations: The Importance of Propertyby-Property Acquisition. *Second Language Research, 27*(3), 379-407. doi:https://doi.org/10.1177/0267658319884116

The acoustic characteristics of voiced bilabials in Basaá: A three-way contrast of implosives, explosives and nasal elements

William Bordonaro[1], Titia Benders[2], Silke Hamann[3], Rodolphe Maah[4]
[1,2,3]University of Amsterdam, [4]University of Douala

Implosive consonants are characterised by substantial cross-linguistic acoustic variation. Although traditionally characterised by ingressive airflow and laryngeal lowering, they can occur without the former and share the latter with voiced plosives (Clements et al., 2002). Their variability suggests that they are abstract typological categories realised through various combinations of articulatory movements, airflow directions, timing and phonation types (Lindau, 1984). Previous research examined implosive-explosive contrasts, finding that implosives may exhibit shorter negative VOT than explosives, as well as higher amplitude and F0 during oral closure and at vowel onset (Mori, 2023). This exploratory study extends this line of research by investigating voiced (naso-)oral closure duration, intensity and F0 in a three-way bilabial contrast between the voiced implosive /ɓ/, the prenasalised explosive voiced stop /ᵐb/, and the syllabic nasal - implosive sequence /m̩ɓ/ in Basaá, a Northwest Bantu language spoken in Cameroon (Makasso & Lee, 2015).

The analysis included 144 word-tokens uttered by four adult L1 Basaá speakers (two male and two female) from the village of Ndjockloumbè (Sanaga-Maritime, Littoral Province), comprising nine word-items, with three items for each consonant category and four repetitions per speaker. The data was manually annotated on Praat TextGrids (Boersma & Weenink, 2019) and extracted with a custom-made Praat script. Duration was calculated from the onset of voiced oral or naso-oral closure to the consonant burst release. F0 and intensity were sampled at 21 intervals of 10 ms across a 200-ms window centred on the consonant burst. Because intensity is a relative measure, values were normalised by subtracting the intensity value of the 21st sampling interval (+100 ms into the following vowel) from all other intervals. Given the nature of the study and the small number of speakers, the results were analysed descriptively. Further statistical analyses are currently underway to assess the significance of the observed patterns.

As shown in Figure 1, the three consonant categories differed clearly in terms of closure duration. Across the combined data comprising all speakers, mean closure duration was shortest for /ɓ/ (77.42 ms, SD = 34.99), followed by /ᵐb/ (135.48 ms, SD = 31.16), while /m̩ɓ/ showed the longest closure duration (211.77 ms, SD = 46.26). The same ordering was also observed individually in each speaker, suggesting that a preceding nasal element contributes substantially to longer duration. Further, the considerably longer closure duration of /m̩ɓ/ may additionally reflect the syllabicity of the nasal element /m̩/, which functions as a tone-bearing unit in Basaá, distinguishing the syllabic nasal-implosive sequence from the prenasalised stop /ᵐb/.

In terms of intensity, the highest pre-release values were observed for /m̩ɓ/, followed by /ɓ/ and /ᵐb/, as reported in Figure 2. The relatively high intensity associated with /m̩ɓ/ may reflect sustained voicing during naso-oral closure. F0, on the other hand, provided a less consistent distinction between the consonant categories: the highest mean F0 at release was observed for /ɓ/, but the overall F0 trajectories varied across speakers. Instances of item-specific variation may also be associated with vowel height and tone, although these factors are not independently balanced in the present materials and therefore cannot be separated from consonant-category effects.

These findings contribute to the acoustic documentation of Basaá and to the cross-linguistic understanding of how implosives are realised within contrastive sets. In particular, the three-way comparison indicates that duration may provide a clearer distinction between /ɓ/ , /ᵐb/ and /m̩ɓ/ than F0 or intensity in the present data. Future research with a larger speaker sample, balanced tonal and vowel-quality conditions, and inferential modelling will be necessary to determine the contribution of these cues to the acoustic description of Basaá.

Figure 1. *Duration of the contrastive consonant categories. /m.b/ = /m̩b/*

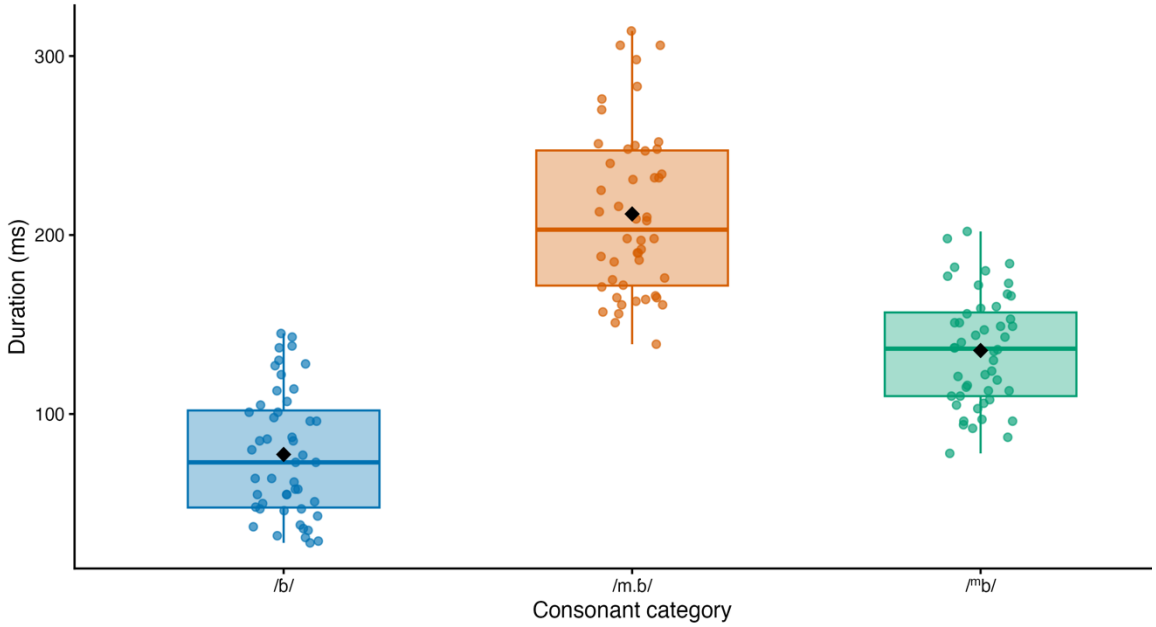
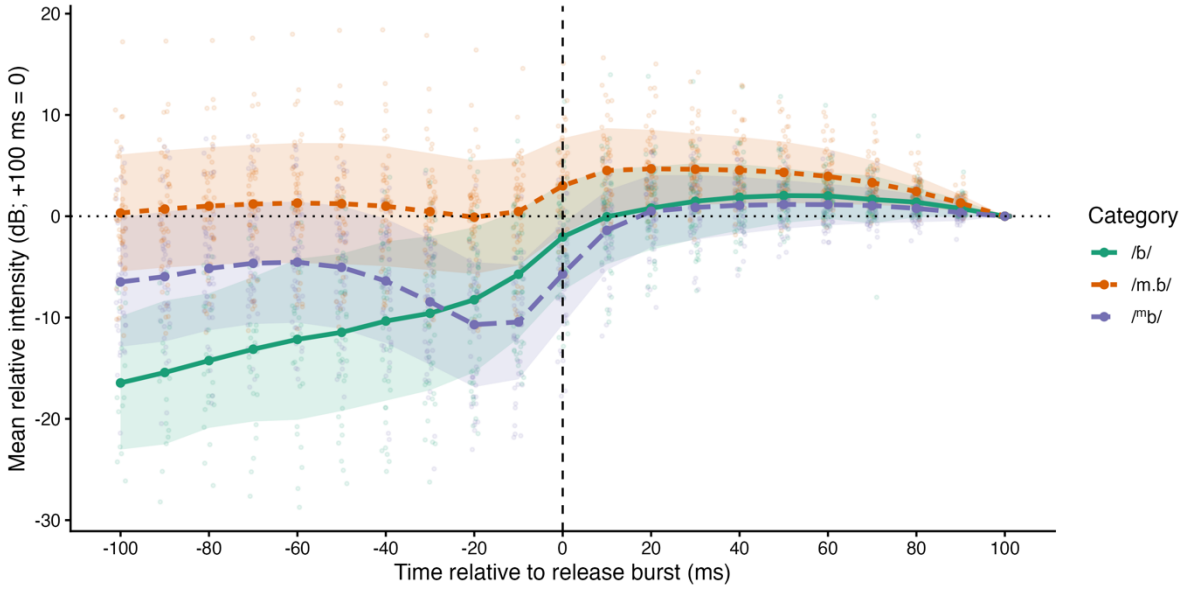


Figure 2. *Mean intensity trajectories by contrastive consonant, with ±1 SD and individual observations (faint points). /m.b/ = /m̩b/*



References

Boersma, Paul & David Weenink. 2019. Praat: Doing phonetics by computer (version 6.6.30). http://www.praat.org

Clements, G. N., Osu, S., Gussenhoven, C., & Warner, N. (2002). Explosives, implosives and nonexplosives: The linguistic function of air pressure differences in stops (pp. 299-350). Berlin & New York: Mouton de Gruyter.

Lindau, M. (1984). Phonetic differences in glottalic consonants. *Journal of Phonetics*, *12*(2), 147–155.

Makasso, E. M., & Lee, S. J. (2015). Basaá. *Journal of the International Phonetic Association*, *45*(1), 71-79.

Mori, M. (2023). The acoustic characteristics of implosive and plosive bilabials in Shimaore. *Journal of the International Phonetic Association*, *53*(3), 950-976.

Vowel reduction based on prosodic position in Dutch

Tristan Bouverot[1]
[1]Universiteit Amsterdam

Vowel reduction in Dutch is a phenomenon which only targets vowels found within unstressed syllables (Kager, 1989). However, not all unstressed vowels are equal in terms of reduction. There is a multitude of factors contributing to vowel reduction in Dutch, including, but not limited to: Word position, syllable weight, segmental context, sociolinguistic context, vowel quality, and word-internal prosody.

For instance, Dutch vowel reduction can occur only word internally, preferably in open syllables beginning in a consonant other than /ɦ/ (Booij, 1995; van Oostendorp, 2000). Vowel reduction tends to only occur in less formal contexts, as it is considered sub-standard (van Oostendorp, 2000; Booij, 1995; Kager, 1989). Specific vowel quality also matters in vowel reduction in the sense that some vowels are less often reduced than others. Distinct vowel reducibility hierarchies have been proposed in Booij (1995) and Kager (1989), based on vowel quality. Vowels at the top of the hierarchy (i.e. /e/) are more reduced than vowels at the bottom of the hierarchy (i.e. /u/). Despite the lengths at which vowel reduction has been discussed in Dutch, conclusions have been almost exclusively made based on impressionistic data. The only instrumental study of Dutch vowel reduction is to be found in van Bergem (1994).

The variable I focus on the most here is word-internal prosody, as some prosodic positions within the word are inherently more prominent than other ones, even among unstressed syllables. For instance, in sequences of syllables embedded between two stressed syllables, the second syllable is usually less reduced than the first one (Kager, 1989). Consider Dutch *locomotief* ($_{,}\sigma\sigma_1\sigma_2{}'\sigma$). In this example, σ_1 tends to be reduced more than σ_2. Multiple theories have been proposed in order to account for such differences of prominence between word-internal unstressed syllables, usually by recourse to prosodic structures intermediate to the syllable and the word.

The most convincing account of differential vowel reduction in unstressed syllables is provided in Kager & Martínez-Paricio (2018), by positing an internally-layered foot structure in Dutch (Martínez-Paricio, 2012; Martínez-Paricio & Kager, 2015). Following this account, foot initial positions are posited as inherently more prominent than other positions within the foot, and feet in Dutch can be recursive (e.g. $(\sigma_1('\sigma_2\sigma_3)_{FtMin})_{FtMax}$). For instance, within $(\sigma_1('\sigma_2\sigma_3)_{FtMin})_{FtMax}$, σ_2 is more prominent than σ_1, because it bears stress, which is more prominent than σ_3, because it is foot initial, hence the fact that σ_1 more readily reduces than σ_3. In the case of Dutch *locomotief* ($_{,}\sigma_1\sigma_2\sigma_3{}'\sigma_4$), Kager and Martínez-Paricio (2018) argue that the difference in reduction behavior between σ_2 and σ_3 is due to the internally layered foot structure of the word: ($_{,}\sigma_1\sigma_2)(\sigma_3('\sigma_4)$)

The main goal of the current study is to acoustically measure whether vowels contained within series of two syllables embedded between two stressed syllables are indeed reduced to different degrees in Dutch, as this has never been done before. If there is an significant observable difference, I will additionally attempt to verify whether Kager & Martínez-Paricio (2018)'s internally layered foot account correctly maps onto the collected data.

In order to collect data, a production task will be conducted. Participants will be recorded reading Dutch sentences containing words of four to five syllables containing identical unstressed vowels in different prosodic contexts. The prosodic contexts are unstressed strong (i.e. foot-initial) and unstressed weak (i.e. non-foot-initial). The list has already been devised through queries on the CELEX corpus (Baayen et al., 2022).

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (2022). CELEX2. https://doi.org/10.5683/SP2/XGW4WY

Booij, G. E. (1995). The phonology of Dutch (1. publ. in paperback). Oxford Univ. Press

Kager, R. (1989). A Metrical Theory of Stress and Destressing in English and Dutch [Doctoral dissertation, Universiteit Utrecht].

Kager, R., & Martínez-Paricio, V. (2018). The internally layered foot in Dutch. Linguistics, 56(1), 69–114. https://doi.org/10.1515/ling-2017-0037

Martínez-Paricio, V. (2012). Superfeet as Recursion. In *Proceedings of the 30th West Coast Conference on Formal Linguistics*. Cascadilla Proceedings Project, Somerville, MA.

Martínez-Paricio, V., & Kager, R. (2015). The binary-to-ternary rhythmic continuum in stress typology: Layered feet and non-intervention constraints. Phonology, 32(3), 459–504. https://doi.org/10.1017/S0952675715000287

van Bergem, D. R. (1994). A model of coarticulatory effects on the schwa. Speech Communication, 14(2), 143–162. https://doi.org/10.1016/0167-6393(94)90005-1

van Oostendorp, M. (2012). Quantity and the Three-Syllable Window in Dutch Word Stress. Language and Linguistics Compass, 6(6), 343–358. https://doi.org/10.1002/lnc3.339

A semi-spontaneous Dutch speech dataset for speech enhancement and speech recognition

Dimme de Groot, Yuanyuan Zhang, Jorge Martinez, Odette Scharenborg
Delft University of Technology

Automatic speech recognition (ASR) systems are widely used in diverse scenarios, including emergency centres, health care services, and education. To ensure reliable performance in practice, it is crucial that ASR systems are able to recognize speech from diverse speakers and under diverse acoustic conditions, including noisy environments with background talkers. However, most speech datasets for developing or evaluating noise-robust ASR and speech enhancement (SE) algorithms consist of synthetic noisy speech, created by mixing clean recordings with real or artificially generated noise. Despite their obvious usefulness, such synthetic noisy speech does not fully capture the complex acoustic characteristics of real noisy speech.

Most publicly available Dutch speech datasets, such as the Corpus Gesproken Nederlands (CGN), Jasmin-CGN, and the Dutch portion of Common Voice 12, were recorded in quiet environments, occasionally with a few noisy samples. Consequently, the performance of state-of-the-art (SotA) ASR systems on Dutch is typically only evaluated on clean speech. For example, OpenAI's Whisper models were evaluated on Dutch speech from Jasmin-CGN and Common Voice 12. Meta's Wav2vec 2.0 model was evaluated on Jasmin-CGN, while two Dutch ASR models trained from scratch were evaluated on Jasmin-CGN.

To answer the open question of how well SotA ASR systems perform on real noisy Dutch speech, we designed and collected DRES. DRES is a freely available 1.5-hour Dutch realistic elicited (semi-spontaneous) speech dataset from 80 speakers recorded in noisy, public indoor environments with background talkers. The speech was recorded with a four-channel linear microphone array and was elicited by requesting the speakers to complete two tasks:

1. Draw a picture card and tell a short story fitting the drawn card,
2. Draw a topic card and shortly talk about that topic.

Using DRES, we evaluated the ASR performance of eight SotA off-the-shelf ASR models developed by Google, Microsoft, Meta, OpenAI, NVIDIA, and the authors. We found that five out of the eight ASR models achieve word error rates (WERs) lower than 22% on DRES, despite the challenging conditions. Additionally, we also evaluated the speech quality and ASR performance after applying six well-known single-channel SE algorithms. In contrast to recent work, we did not find a positive effect of modern single-channel SE on ASR performance, emphasizing the importance of evaluating in realistic conditions.

This work has been accepted for presentation at IEEE SLT 2026.

References

Zhang, Y., De Valck, T., & Scharenborg, O. (2026). Speech recognition performance disparities between Dutch diverse speaker groups. Phonetica, (0). https://doi-org.tudelft.idm.oclc.org/10.1515/phon-2025-0061

Fuckner, M., Horsman, S., Wiggers, P., & Janssen, I. (2023, October). Uncovering bias in ASR systems: Evaluating wav2vec2 and whisper for Dutch speakers. In *2023 International conference on speech technology and human-computer dialogue (SpeD)* (pp. 146-151). IEEE. 10.1109/SpeD59241.2023.10314895

Talking to a brick wall? Examining the role of a priori assumptions in computer-directed speech

Jurjen de Haan[1], Khiet P. Truong[1 2]
[1]Human Media Interaction, University of Twente, [2]Centre for Language Studies, Radboud University

Several studies have shown that speakers do not address computers in the same way as humans, leading to the identification of computer-directed speech as a distinct speech register (e.g., (Burnham et al., 2010; Cohn et al., 2022)). Findings of slower speaking rates (Cohn et al., 2024), increased loudness (Raveh et al., 2019), and vowel hyperarticulation (Burnham et al., 2010) resemble patterns of clear speech (Smiljanic, 2021; Smiljanić & Bradlow, 2009), and suggested reasons for adaptations commonly focus on speakers' mental models of computers as having limited communicative competence (Cohn et al., 2024; Cohn et al., 2022; Ibrahim & Skantze, 2021; Raveh et al., 2019; Zellou & Holliday, 2024).

However, there is a lack of evidence that speech adaptations toward computers are systematic patterns of variation associated with addressing a computer rather than a human or other broad category of interlocutor, as the term "computer-directed speech register" implies. Human and computer conditions in previous studies do not only differ in interlocutor category, but also in interlocutor-specific factors such as error rate and voice type, which may drive differences in speech production between conditions through hyperarticulation (Lindblom, 1990; Smiljanic, 2021) and phonetic alignment (Pickering & Garrod, 2004), respectively.

To address this problem, this study aimed to isolate the effect of participant's belief of their interlocutor's identity on speech production using a controlled experiment. Participants in the computer condition were instructed to perform an interactive task with a computer while those in the human condition were instructed to perform the task with another person (between-subjects factor). Human and computer conditions were controlled by the researcher in a Wizard-of-Oz style setup and were near-identical[1] in perceptual and interactional cues throughout the task, to control for the effects of interlocutor-specific factors on speech production. The task structure yielded original productions, confirmations, and error-repair utterances for analysis (within-subjects factor). The speech variables examined in this study include speaking rate, mean f0, f0 range, vowel space area, and vowel duration.

Beyond the effects of interlocutor identity and utterance type on speech production, this study additionally investigated whether potential speech differences in human and computer conditions are specifically driven by speakers' expectations of their interlocutor's ability to comprehend spoken language. To do this, participant ratings of how well they expect their interlocutor to understand them were collected before the task and were assessed as predictors of the speech variables.

Contrary to the hypothesis, interlocutor identity did not have an effect on overall speech style. Instead, speech differences were driven by local communicative demands, including repairing misunderstandings and confirming sentences that were correctly understood. Participants' expectations of their interlocutor's ability to understand them were not significantly associated with the speech variables. The findings contribute to our understanding of the sources of computer-directed speech and motivate further research on the relative contributions of a priori assumptions, perceptual differences, and local communicative demands.

[1] Human and computer conditions differed in a name presented on the task screen throughout the experiment. Otherwise, the conditions were perceptually and interactionally identical.

References

Burnham, D., Joeffry, S., & Rice, L. (2010). Computer- and human-directed speech before and after correction. In *Australasian International Conference on Speech Science and Technology* (pp. 13-17): ASSTA.

Cohn, M., Barreda, S., Graf Estes, K., Yu, Z., & Zellou, G. (2024). Children and adults produce distinct technology- and human-directed speech. *Scientific Reports*, *14*(1), 15611. https://doi.org/10.1038/s41598-024-66313-5

Cohn, M., Segedin, B. F., & Zellou, G. (2022). Acoustic-phonetic properties of Siri- and human-directed speech. *Journal of Phonetics*, *90*, 101123. https://doi.org/https://doi.org/10.1016/j.wocn.2021.101123

Ibrahim, O., & Skantze, G. (2021, 2021). *Revisiting robot directed speech effects in spontaneous Human-Human-Robot interactions* Human Perspectives on Spoken Human-Machine Interaction, https://doi.org/10.6094/UNIFR/223822http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-312486

Lindblom, B. (1990). Explaining Phonetic Variation: A Sketch of the H&H Theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403-439). Springer Netherlands. https://doi.org/10.1007/978-94-009-2037-8_16

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169-190. https://doi.org/10.1017/S0140525X04000056

Raveh, E., Steiner, I., Siegert, I., Gessinger, I., & Möbius, B. (2019, 2019/3). *Comparing phonetic changes in computer-directed and human-directed speech* Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019, https://www.essv.de/pdf/2019_42_49.pdf

Smiljanic, R. (2021). Clear Speech Perception. In *The Handbook of Speech Perception* (pp. 177-205). https://doi.org/https://doi.org/10.1002/9781119184096.ch7

Smiljanić, R., & Bradlow, A. R. (2009). Speaking and Hearing Clearly: Talker and Listener Factors in Speaking Style Changes. *Language and Linguistics Compass*, *3*(1), 236-264. https://doi.org/https://doi.org/10.1111/j.1749-818X.2008.00112.x

Zellou, G., & Holliday, N. (2024). Linguistic analysis of human-computer interaction [Review]. *Frontiers in Computer Science*, *Volume 6 - 2024*. https://doi.org/10.3389/fcomp.2024.1384252

The Role of Social Network Size and Diversity in Vowel Identification in Noise

Sarah Kapp[1], Adriana Hanulíková[1]
[1]University of Heidelberg

Individuals differ in the variability of speech input they receive in everyday life. Research shows that this variability affects language acquisition and linguistic skills and has recently been operationalized using properties of individuals' social networks (i.e., the people they regularly interact with) such as their size and diversity (Hanulíková & Levy, 2025; Kutlu et al., 2024; Lev-Ari, 2016, 2018; Levy et al., 2019). Network *size* refers to the number of people an individual interacts with, whereas *diversity* refers to the degree of linguistic heterogeneity among them. Previous studies have mainly focused on network size as a proxy for input variability. Although this measure reflects input variability to some extent (Lev-Ari, 2016, 2018), it provides a rather rough estimate, as it does not account for the linguistic characteristics of network members' speech. The present study addresses this limitation by investigating whether linguistic diversity within social networks provides additional information about speech perception beyond network size.

The few studies including measurements of linguistic diversity assess characteristics known to contribute to systematic pronunciation variation, including regional and foreign accents as well as exposure to different phonological systems through other languages (Cowan & Olmstead, 2023; Hanulíková & Levy, 2025; Kutlu et al., 2024; Levy et al., 2019). Beneficial effects of input variability by exposure to multiple speakers or linguistic diversity have been found for phonological processing (Bradlow & Bent, 2008; Sumner, 2011) and speech comprehension (Hanulíková, 2021; Levy et al., 2019) and are often explained by the notion that variable input leads to the formation of robust categories (Raviv et al., 2022). Lev-Ari (2018) found that participants with larger social networks performed better in a vowel perception task. Expanding this research, we investigate whether German vowel identification in noise is predicted by social network size and, beyond that, by the linguistic diversity within these networks. We expected larger and more linguistically diverse networks to be associated with better vowel identification performance, reflecting additional benefits of exposure to linguistically variable speech input.

Participants (n=61) completed a questionnaire on their social networks, a vowel-identification-in-noise task, and two control tasks assessing attention skills (Flanker task) and receptive vocabulary (LexTALE). The main task included 126 monosyllabic nonwords, embedded in speech-shapes noise, across which 14 vowels of Standard German were evenly distributed. Participants responded with the perceived vowel. Network size was operationalized as the number of individual network members and diversity as the number of network members who spoke with a regional or foreign accent or in another language. Additionally, three separate diversity scores for regional accents, foreign accents and other languages were calculated, reflecting the percentage of the total weekly speech input associated with each variety, as effects of one type of input variability do not necessarily generalize to others (Levy et al., 2019).

Average identification accuracy was 59.46% (SD=11.67) and network size ranged from 2 to 28 members (M=9.58). Results model showed that participants with larger networks showed higher identification accuracy, suggesting that real-life exposure to many speakers may benefit speech perception and supporting the idea that this exposure may contribute to more robust phonological representations. Higher attention skills were also associated with better performance, indicating that attention skills may contribute to a challenging listening task or that participants with better attention skills may have been better able to maintain focus during the task. The lack of an effect of diversity could suggest that the measures used did not sufficiently capture the aspects of variability that are relevant for vowel identification or that the variability reflected by these measures may support other linguistic skills but not the specific skill testes in this study.

References

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Cowan, T. M., & Olmstead, A. J. (2023). Exploring the Implications of Input Variability for Unfamiliar Accented Speech Perception: A Focused Review and New Hypotheses. *Languages*, *8*(1), 67. https://doi.org/10.3390/languages8010067

Hanulíková, A. (2021). Do faces speak volumes? Social expectations in speech comprehension and evaluation across three age groups. *PLOS ONE*, *16*(10), e0259230. https://doi.org/10.1371/journal.pone.0259230

Hanulíková, A., & Levy, H. (2025). Quantifying Experience with Accented Speech to Study Monolingual and Bilingual School-Aged Children's Speech Processing. *Languages*, *10*(4), Article 4. https://doi.org/10.3390/languages10040080

Kutlu, E., Baxelbaum, K., Sorensen, E., Oleson, J., & McMurray, B. (2024). Linguistic diversity shapes flexible speech perception in school age children. *Scientific Reports*, *14*(1), 28825. https://doi.org/10.1038/s41598-024-80430-1

Lev-Ari, S. (2016). How the Size of Our Social Network Influences Our Semantic Skills. *Cognitive Science*, *40*(8), 2050–2064. https://doi.org/10.1111/cogs.12317

Lev-Ari, S. (2018). The influence of social network size on speech perception. *Quarterly Journal of Experimental Psychology*, *71*(10), 2249–2260. https://doi.org/10.1177/1747021817739865

Levy, H., Konieczny, L., & Hanulíková, A. (2019). Processing of unfamiliar accents in monolingual and bilingual children: Effects of type and amount of accent experience. *Journal of Child Language*, *46*(2), 368–392. https://doi.org/10.1017/S030500091800051X

McQueen, J. M. (2005). Speech perception. In K. Lamberts & R. Goldstone (Eds.), *The handbook of cognition* (pp. 255–275). SAGE. https://www.torrossa.com/gs/resourceProxy?an=4912059&publisher=FZ7200#page=276

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*, *26*(6), 462–483. https://doi.org/10.1016/j.tics.2022.03.007

Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition*, *119*(1), 131–136. https://doi.org/10.1016/j.cognition.2010.10.018

Hearing smiles in the crowd: how babble noise shapes smiled speech perception

Rong Li[1], Esther Janse[2], Dirk Heylen[1], Khiet P. Truong[2,1]
[1]Human Media Interaction, University of Twente, the Netherlands,
[2]Centre for Language Studies, Radboud University, the Netherlands

Smiles are ubiquitous in everyday interaction and are conveyed not only visually but also through the voice, where they are reflected in systematic variations in prosody and voice quality. For instance, smiling can increase the second formant (F2) frequency, particularly in rounded vowels, and has been associated with increased intensity and changes in higher-formant energy (Barthel & Quené, 2015; Ponsot et al., 2018). Listeners can also detect smiles in speech from voice alone. In particular, spread-lip speech primarily reflects the acoustic consequences of the facial gesture, whereas amused speech additionally involves prosodic modulation beyond these mechanical effects (Aubergé & Cathiard, 2003; Drahota et al., 2008). Yet auditory smile perception has largely been studied under clean laboratory listening conditions. We therefore investigated whether smile-related vocal information remains robust under multi-talker babble noise.

We examined two aspects of auditory smile perception: (1) detecting a smile-like quality in speech and (2) categorizing smile-like speech as amused or spread-lip. Stimuli were selected from the AMuS corpus (El Haddad et al., 2017) and comprised neutral, amused, and spread-lip productions from one French-speaking male and one English-speaking female (with the French materials presented to French-fluent and English materials to English-fluent listeners). Recordings were presented in three listening conditions: Quiet, −3 dB, and −6 dB SNR, using multi-talker cafeteria babble from the NOISEX-92 database (Varga & Steeneken, 1993).

Seventy-five adults (38 English-fluent and 37 French-fluent listeners) completed two online binary forced-choice tasks. Task 1 contrasted smile-like speech (amused and spread-lip) with neutral speech, whereas Task 2 contrasted amused with spread-lip speech. Sentence materials were rotated across noise conditions so that each sentence occurred only once per participant. Trial-level classification accuracy was analyzed using generalized linear mixed-effects models (GLMMs), complemented by Signal Detection Theory (SDT) analyses (Hautus et al., 2021) of perceptual sensitivity (d'), decision criterion (c), and Area Under the ROC Curve (AUC).

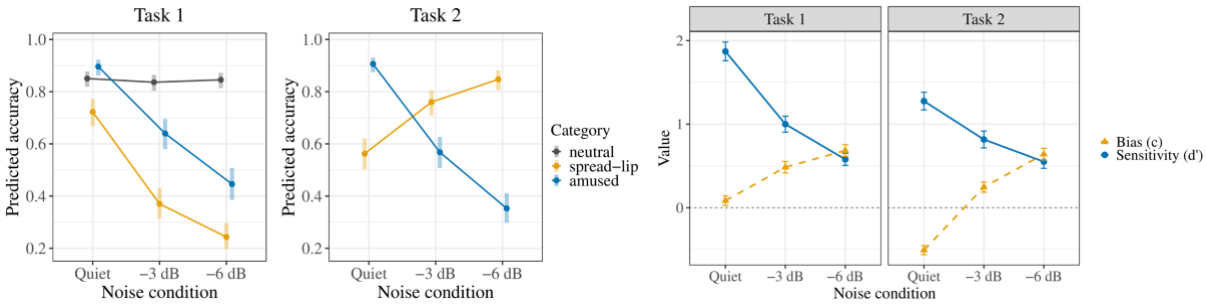
In Task 1, classification accuracy for both spread-lip and amused speech deteriorated substantially with increasing noise, whereas neutral-speech detection remained relatively stable, yielding a significant Noise x Category interaction. Amused speech was recognized more accurately than spread-lip speech across all noise levels. Perceptual sensitivity decreased systematically with increasing noise, accompanied by a progressive shift toward a more conservative decision criterion (Figure 1). Thus, participants not only had less sensory evidence available for smile judgments but also became less inclined to report hearing a smile.

Task 2 showed a pronounced interaction between noise condition and smile category. Under quiet conditions, amused speech was classified substantially more accurately than spread-lip speech. Under increasing noise, amused accuracy declined while spread-lip accuracy increased, producing a crossover pattern (Figure 1). Sensitivity for distinguishing the two smile types nevertheless decreased steadily, while a liberal bias toward "amused" responses in quiet reversed to a conservative bias under noise.

Across both tasks, noise masking therefore reduced the sensory evidence available for smile judgments while also shifting listeners toward a more conservative decision strategy as uncertainty increased. The findings highlight the importance of separating perceptual sensitivity from response bias when studying smiled speech in realistic noisy environments and when evaluating speech technologies operating without visual context.

Figure 1

Effects of noise on auditory smile perception



Note. Left: GLMM-predicted accuracy by task, category, and noise condition (95% CIs). Right: mean sensitivity (d') and decision criterion (c) across noise conditions for both tasks (SE). Reproduced from Li et al. (2026).

Author Note

This work has been published in the *Proceedings of Interspeech 2026* (Li et al., 2026). The present abstract summarizes the same study and results.

References

Aubergé, V., & Cathiard, M. (2003). Can we hear the prosody of smile? *Speech Communication, 40*(1), 87–97.

Barthel, H., & Quené, H. (2015). Acoustic-phonetic properties of smiling revised: Measurements on a natural video corpus. In *Proceedings of the 18th International Congress of Phonetic Sciences*.

Drahota, A., Costall, A., & Reddy, V. (2008). The vocal communication of different kinds of smile. *Speech Communication, 50*, 278–287.

El Haddad, K., Torre, I., Gilmartin, E., Çakmak, H., Dupont, S., Dutoit, T., & Campbell, N. (2017). Introducing AmuS: The Amused Speech Database. In *Statistical Language and Speech Processing* (pp. 229–240). Springer International Publishing.

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection theory: A user's guide*. Routledge.

Li, R., Janse, E., Heylen, D., & Truong, K. P. (2026). Hearing smiles in the crowd: How babble noise shapes smiled speech perception. *Proceedings of Interspeech 2026*, 4690–4694.

Ponsot, E., Arias, P., & Aucouturier, J.-J. (2018). Uncovering mental representations of smiled speech using reverse correlation. *The Journal of the Acoustical Society of America, 143(*1), EL19–EL24.

Varga, A., & Steeneken, H. J. M. (1993). Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication, 12*(3), 247–251.

Why can *tigers* be *pigers* but *babies* not *dabies*?
An eye-tracking study investigating the emergence of coronal underspecification in the developing lexicon of 24-months-old Dutch toddlers

Giusy Santo[1 2], Merel van Goch[1], Paula Fikkert[1]
[1] Radboud University, Nijmegen, The Netherlands, [2] International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands

Mapping sounds onto meaning relies on detecting subtle but meaningful contrasts in the speech signal. As the lexicon grows, purely acoustic processing becomes insufficient for rapid and efficient word access. A central question is, therefore, whether infants transition from an acoustically grounded system to one that relies on abstract phonological representations, in which only marked contrastive features are stored in the mental lexicon.

A key model in this debate is the Featurally Underspecified Lexicon model (FUL; Lahiri & Reetz, 2010), which proposes that some phonemes, such as coronals, are underspecified in the mental lexicon for place of articulation (PoA). As a result, such phonemes would be more tolerant of PoA changes, whereas others, such as labials, are fully specified and, therefore, less tolerant of mispronunciations (MPs). Adult studies consistently support this asymmetry, with coronal MPs hindering word recognition less than labial MPs (e.g., Roberts et al., 2013). Developmental findings, however, are less consistent: some report asymmetries (e.g., Feest & Fikkert, 2015, for Dutch toddlers) while others do not (e.g., Ren et al., 2019, for American toddlers). However, methodological inconsistencies (e.g. audio- vs picture-first paradigm designs), together with differences in the age and native language of the toddlers tested, make it hard to disentangle whether these discrepancies in the results reflect task demands, cross-linguistic differences in the phonological systems, actual developmental changes in the speech processing system, or growing lexicons containing more words with onsets that contrast in PoA.

While most previous MP studies employ a picture-first design, Althaus et al. (2024) employs an innovative audio-first intermodal preferential looking (IPL) paradigm, claiming that whereas a picture-first design may encourage surface-level sound-matching strategies, an audio-first design may more directly tap into toddlers' stored phonological representations. Interestingly, their results in 24- months-olds align with FUL predictions: labial-to-coronal MPs (e.g., *[daby] for /baby/) hindered lexical access more than coronal-to-labial MPs (e.g., *[piger] for /tiger/). No asymmetry, however, was observed in the younger group of 18-month-olds. This methodological innovation is particularly relevant for Dutch, where Fikkert and colleagues have consistently reported coronal underspecification from as early as 4 months through the second year of life (Dijkstra & Fikkert, 2011; Altvater-Mackensen et al., 2014; Tsuji et al., 2015, 2016; Feest & Fikkert, 2015) – but always using picture-first paradigms.

The present study addresses this methodological gap by replicating the audio-first IPL paradigm used by Althaus et al. (2024) with 24-month-old Dutch toddlers, testing whether this asymmetry generalizes cross-linguistically within the same age range. Participants heard either correct or mispronounced labels (8 coronal-onset, 8 labial-onset target words) before viewing paired target and distractor images (see Figure 1 for trial timeline). Target-looking proportions from image onset to offset (3000 ms window) will be analysed using growth curve analysis, comparing coronal vs labial MPs over time. Following FUL, we predict reduced target-looking proportions following labial-to-coronal as compared to coronal-to-labial MPs. In addition, to investigate whether lexicon size plays a role, we examine individual differences in MP sensitivity in relation to (theta) values of lexical development obtained from the NL-CDI:CAT (Sametoğlu et al., 2025).
Data collection is currently ongoing (approximately 40% of the target sample tested to date) and is expected to be completed by the time of the event, when preliminary results will be presented.

Figure 1. Trial timeline.

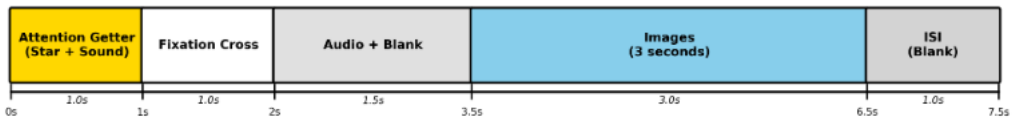


Figure 2. Examples of trials with target and distractor images.





References

Althaus, N., Lahiri, A., & Plunkett, K. (2024). Coronal underspecification as an emerging property in the development of speech processing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 50(12), 1932.

Altvater-Mackensen, N., van der Feest, S. V., & Fikkert, P. (2014). Asymmetries in early word recognition: The case of stops and fricatives. Language Learning and Development, 10(2), 149-178.

Lahiri, A., & Reetz, H. (2010). Distinctive features: Phonological underspecification in representation and processing. Journal of Phonetics, 38(1), 44-59.

Ren, J., Priva, U. C., & Morgan, J. L. (2019). Underspecification in toddlers' and adults' lexical representations. Cognition, 193, 103991.

Roberts, A. C., Wetterlin, A., & Lahiri, A. (2013). Aligning mispronounced words to meaning: Evidence from ERP and reaction time studies. The Mental Lexicon, 8(2), 140-163.

Sametoğlu, S., Matetovici, M., Van den Akker, M., & Rowland, C. F. (2025). Online computerized adaptive tests of children's vocabulary development in Dutch. Talk presented at the MEDAL Conference: New Challenges, Novel Approaches. Tartu, Estonia. 2025-10-09 - 2025-10-10.

Tsuji, S., Fikkert, P., Yamane, N., & Mazuka, R. (2016). Language-general biases and language-specific experience contribute to phonological detail in toddlers' word representations. Developmental psychology, 52(3), 379.

Tsuji, S., Mazuka, R., Cristia, A., & Fikkert, P. (2015). Even at 4 months, a labial is a good enough coronal, but not vice versa. Cognition, 134, 252-256.

Van der Feest, S. V., & Fikkert, P. (2015). Building phonological lexical representations. Phonology, 32(2), 207-239.

Acknowledgements

This work was funded by the ERC Synergy grant (101167219) awarded to Paula Fikkert, Linda Wheeldon, Allison Wetterlin and Aditi Lahiri.

LLMs at the Phonetics-Phonology-Orthography Interface

Zoe Tholen
University of Amsterdam

The popularization of Large Language Models (LLMs) has sparked discourse both about if an LLM could ever attain "true" knowledge, and about how this knowledge could be utilized to aid the sciences. The current study followed the behavioral testing procedures suggested by Ribeiro et al. (2020) to see if Google's Gemini Flash and OpenAi's GPT-5.1 employ human-like (productive, layered) phonetic and phonological "knowledge" while transcribing into and from IPA, as well as English and Korean orthography, or if they merely replicate IPA transcriptions from their training dataset. In the first test, models were asked to transcribe 14 English pseudowords (presented in IPA) into English orthography. The second test assessed whether the same pseudowords in English orthography could be implemented into Korean orthography, while retaining English phonotactics, rather than adapting Korean phonotactics. Lastly, a third test introduced an additional syllable (occurring in both languages) to the second test's stimuli and checked whether this needlessly affected the rest of the transcription. The tests were evaluated for the LLMs' abilities to recognize and rank language-specific phonotactic demands to implement phonetic information into the requested orthographical representations. For example, when English loanwords are adapted into Korean phonotactics, they sometimes experience syllabification as this allows for the maintenance of finer phonetic details in coda plosives (e.g.: notebook → 노트북 [no.tʰɯ.buk]), whereas a native English speaker might prefer to sacrifice phonetic detail in order to maintain the correct syllable structure (e.g.: notebook → 놑북 [not̚.buk]). The second test assessed whether such a non-canonical, English-favouring transcription could be applied by the LLMs.

Generally, the models performed well on all tasks, with failure rates between about 7% and 29% across models and tests. However, both LLMs scored worse on the third than on the second test, despite the simplicity of the additional task, perhaps indicating that the LLMs did not productively apply phonological rules in a human-like manner. The specific mistakes made and explanations offered by the LLMs further suggested that they underprioritized language-specific phonology and inappropriately mimicked input from their training datasets. These findings imply that publicly available LLMs may mostly be able to imitate productive phonological knowledge, but might not readily be used as aids in research demanding phonetic and phonological transcription or analysis without extensive prompting and human oversight.

References

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. *Annual Meeting of the Association for Computational Linguistics*, 4902–4912. https://doi.org/10.18653/v1/2020.acl-main.442

Grammar or grammars? A maximum entropy analysis of loanword adaptation in Shanghainese

Meng Zhang
Utrecht University

This study investigates to what extent a computational model could capture segmental patterns of loanword adaptation in Shanghainese, a Wu-dialect of Mandarin Chinese (Chen & Gussenhoven, 2015), under the framework of Optimality Theory (OT; Prince & Smolensky, 1993). It is anticipated that one overarching grammar would be sufficient to account for the whole adaptation process.

While there are substantial OT studies in the realm of Sinitic loanword phonology, a trade-off persists between the scale of the data used and the ability of generalization from the findings. Moreover, Shanghainese sound inventories and loanword lexicon exhibit unique yet understudied lexical items and phonemic features such as the prevalence of voiced phonemes that could bring new insights to the matter.

Using a self-compiled corpus of 197 loanwords, four computational models were built and compared under Maximum Entropy analysis (MaxEnt; Goldwater & Johnson, 2003), a probabilistic extension of OT. The four models differed in whether consonants only or all segments are considered, and whether the learning process was carried out (lexical) item by item or for the entire corpus

The findings, however, challenge the assumption. While per-item models have achieved almost perfect performance at the ceiling, global models, in sharp contrast, displayed moderate to poor learnability, capturing only general or almost no pattern at all. Following this, diagnostics on both the constraints and dataset were conducted to seek out possible explanations.

Together, the results suggest that Shanghainese loanword adaptation is perhaps better characterized by a compilation of sub-grammars, rather than a single, unified ranking of constraints. Apart from implications of segmental asymmetry and structured variation, this study also discusses the empirical limitation of MaxEnt and OT-style analysis, as well as influential factors of loanword adaptation beyond sound changes.

References

Chen, Y., & Gussenhoven, C. (2015). Shanghai Chinese. *Journal of the International Phonetic Association, 45*(3), 321–337. https://www.jstor.org/stable/26352263

Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Workshop on Variation within Optimality Theory* (pp. 111 – 120). http://homepages.inf.ed.ac.uk/sgwater/papers/OTvar03.pdf

Prince, A., & Smolensky, P. (1993). *Optimality Theory: Constraint interaction in generative grammar* (Technical Report No. RuCCS-TR-2). Rutgers University Center for Cognitive Science. http://roa.rutgers.edu/view.php3?roa=537