Middag van de Fonetiek

ABSTRACTS

18 december 2020

Online

PROGRAMMA

keynote	12:30-13:15	About good and bad prosody	Marc Swerts
pauze	13:15-13:30	*BreakOut rooms open*	
1	13:30-13:50	Automatic assessment of transcript accuracy for speech intelligibility studies	Hans Rutger Bosker
2	13:50-14.10	Forced Alignment: een krachtig hulpmiddel voor onderzoek aan spraak	Arjan van Hessen & Louis ten Bosch
pauze	14:10-14:20	*BreakOut rooms open*	
3	14:20:14:40	The Perception-Production Link in Learning Words with Lexical Tone	Tim Laméris & Brechtje Post
4	14:40-15:00	Binnensprekervariatie in de uitspraak van /m/ in verschillende talen	Meike de Boer & Willemijn Heeren
pauze	15:00-15:10	*BreakOut rooms open*	
5	15:10-15:30	Contour clustering: a tool for exploring prototypical f0 patterns	Constantijn Kaland
6	15:30-15:50	Automatic Analysis of Speech Prosody in Dutch	Na Hu, Berit Janssen, Judith Hanssen, Carlos Gussenhoven & Aoju Chen
pauze	15:50-16:00	*BreakOut rooms open*	
7	16:00-16:20	Listeners learn and predict talker-specific prosodic cues in speech perception	Giulio G.A. Severijnen, Hans Rutger Bosker, Vitória Piai & James M. McQueen
8	16:20-16:40	Prosodic phrasing of short left-dislocated adverbial adjuncts in Brazilian Portuguese	Tainan Carvalho, Luciani Tenani & Marc Swerts
9	16:40-17:00	And now for something completely different…	Vincent J. van Heuven
9	17:00-17:10	ALV	
slot	17:10-18:00	*BreakOut rooms open*	

KEYNOTE

About good and bad prosody

Marc Swerts
Tilburg University

Not all speakers are equally good. For instance, at a scientific conference where one can witness many different speakers, there typically tend to be presenters who are engaging, whereas others are boring. The difference in speaking style between good and bad speakers may be partly related to differences in the way they supplement their utterances with appropriate prosodic structures. In this talk, I will discuss research we did on the extent to which the goodness of a speaking style depends on both functional and formal properties of prosody. The first part of my talk will zoom in on the extent to which quality differences relate to differences in pitch accent distribution. The second part discusses joint work with Constantijn Kaland on the perceived quality of variation in speech rhythm.

Automatic assessment of transcript accuracy for speech intelligibility studies

Hans Rutger Bosker[1]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

HansRutger.Bosker@mpi.nl

In the field of speech perception, many studies assess the intelligibility of spoken stimuli by means of verbal repetition (‘repeat back what you hear’) or transcription tasks (‘type out what you hear’). The intelligibility of a given stimulus is then often expressed in terms of percentage of words correctly reported from the target stimulus. Yet scoring the participants’ raw transcripts for words correctly identified from the target stimulus is a time-consuming task, and hence resource-intensive. Moreover, there is no consensus on what protocol to use for the human scoring, limiting the reliability of human scores. The present paper evaluates various forms of ‘fuzzy string matching’ between participants’ responses and target sentences as automated metrics of listener transcript accuracy. Fuzzy string matching is identified as a consistent, efficient, and accurate method for automated assessment of listener transcripts, as evidenced by high correlations with human-generated scores (highest r = 0.94) and a strong relationship to acoustic markers of speech intelligibility. Thus, fuzzy string matching provides a practical tool for speech scientists, allowing fast and reliable assessment of listener transcript accuracy in large-scale speech intelligibility studies.

Forced Alignment: een krachtig hulpmiddel voor spraakonderzoek

Arjan van Hessen, Universiteit Twente

Louis ten Bosch, Radboud Universiteit, Nijmegen

In veel onderzoek aan spraak wordt gebruik gemaakt van Forced Alignment. Bij een Forced Alignment wordt de orthografische transcriptie van een bepaald audiofragment "opgelijnd" met dat fragment. Deze oplijning houdt in dat het akoestische begin en einde van elk woord zo precies mogelijk worden gezocht als ankerpunten in de audio. En dit oplijnen geldt niet alleen de woorden maar ook de eventuele stiltes voor, tussen en na de woorden. Als resultaat van de Forced Aligner weet je precies hoe lang woorden en stiltes duren. Deze kennis is van groot belang bij bijvoorbeeld onderzoek naar sprekervariatie, naar uitspraakvariatie, spreektempo, en voor het mogelijk maken van het semi-automatisch doorzoeken van audiobestanden via geschreven queries.

Bij het CLST in Nijmegen is in samenwerking met de Stichting Open Spraaktechnologie een aligner gebouwd waarin niet alleen woorden maar ook de spraakklanken in elk woord worden opgelijnd met een audiofile. De resultaten op woord- en foonniveau komen tegelijkertijd beschikbaar als twee tiers in een Praat textgrid file. Daarnaast is het mogelijk de aligner een eigen woordenboek mee te geven waarin bijvoorbeeld specifieke woorden kunnen worden voorzien van afwijkende uitspraakrealisaties. Dat maakt onderzoek aan uitspraakvarianten mogelijk.

In de presentatie gaan we in op de functionaliteit van de aligner in een aantal realistische toepassingen, en op de design filosofie van de forced alignment webservice.

The Perception-Production Link in Learning Words with Lexical Tone

Tim Laméris and Brechtje Post
Phonetics Lab, University of Cambridge

Although it is commonly agreed that speech acquisition in both perception and production are closely intertwined, performance in one modality may not always mirror performance in another. In this study, we present new evidence for the perception-production link by looking at L2 acquisition of lexical tone. We trained a group of English (n=21) and Mandarin Chinese (n=20) speakers to learn a set of 16 words in a tonal pseudolanguage made up of four segments (/jar/, /jur/, /nɔn/ and /lɔn/) and four lexical tones (rising, falling, mid-level, and low-level). After a two-day training session, subjects were tested on their word identification and word production accuracy to assess word learning in both modalities. Normalised f0 data were obtained to determine tone production accuracy. We also accounted for participants' extralinguistic characteristics, such as musical background and working memory.

We found that accuracy, improvement in their performance, and types of errors in the two modalities were highly correlated. Both in listening and speaking, most word recall errors were purely tonal in nature (i.e. often the words' segmental but not their tonal properties were retained), but Mandarin Chinese participants were much more likely than English participants to confuse level tone contrasts, which do not exist in the Mandarin tone inventory. Crucially, these error patters occurred both in word identification and production, with remarkable similarities between both domains.

This study adds to a currently limited body of work on the perception-production link in second language tone-learning, which has mainly focused on perception and production at the pre-lexical level. We show that the perception-production correlation is largely maintained at the lexical level.

Binnensprekervariatie in de uitspraak van /m/ in verschillende talen

Meike de Boer & Willemijn Heeren
Universiteit Leiden, Leiden University Centre for Linguistics

In forensisch zaakonderzoek komt steeds vaker spraakmateriaal in verschillende talen voor. Dit roept de vraag op of er taalonafhankelijke sprekerspecifieke kenmerken zijn. De bilabiale nasaal /m/ is een van de meest sprekerspecifieke segmenten, wat wordt toegeschreven aan de rigiditeit van de neusholte [1]. Tegelijkertijd is de mondholte ook betrokken bij de productie en heeft de tong daarbij geen vaste positie [2]. Hierdoor is er ruimte voor binnensprekervariatie, die mogelijk taalafhankelijk is. Wij onderzochten in hoeverre de realisatie van /m/ verschilt tussen de eerste (L1) en tweede taal (L2) van meertalige sprekers.

Er zijn monologen gebruikt van 53 vrouwelijke sprekers uit D-LUCEA [3], in hun L1 Nederlands en L2 Engels. De sprekers waren eerstejaarsstudenten van University College Utrecht en hadden een bovengemiddelde beheersing van het Engels. De nasalen werden gesegmenteerd in Praat en geanalyseerd op verschillende akoestische kenmerken.

De resultaten laten zien dat de verschillen tussen de realisaties in de L1 en L2 minimaal zijn. Alleen de tweede nasale formant (N2) liet een taalverschil zien: hoger in de L2 dan in de L1. Sprekers verschilden in de mate waarin ze deze verschuiving vertoonden en voor sommigen ging het resultaat in tegengestelde richting.

Hoewel de gevonden L1−L2 verschillen in de uitspraak van de /m/ klein zijn, lijkt de bilabiale nasaal enigszins taalafhankelijk. De N2 wordt gerelateerd aan de mond- en neusholte [4], wat duidt op een aanpassing in de productie. Rekening houdend met deze aanpassing zou de /m/ bruikbaar kunnen zijn in meertalige forensische sprekervergelijkingen. Vervolgstappen zijn om dit te onderzoeken met sprekerclassificatie.

Referenties

[1] Rose, P. (2002). Forensic speaker identification. In: J. Robertson (Ed.), *Taylor & Francis Forensic Science Series*. London: Taylor & Francis (pp. 125-173).
[2] Su, L., Li, K. -P., & Fu, K. S. (1974). Identification of speakers by use of nasal coarticulation. *The Journal of the Acoustical Society of America*, *56*(6), 1876–1883.
[3] Orr, R., & Quené, H. (2017). D-LUCEA: Curation of the UCU Accent Project data. In: J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries*. Berkeley: Ubiquity Press (pp. 177−190).
[4] Fant, G. (1970). *Acoustic theory of speech production* (2nd ed.). The Hague: Mouton.

Contour clustering: a tool for exploring prototypical f0 patterns

Constantijn Kaland
Institute of Linguistics, University of Cologne

This work presents an automatic data-driven analysis for describing prototypical f0 patterns. This is particularly suitable as an exploratory tool in initial stages of prosodic research and language description. The approach has several advantages over traditional ways to investigate prosody and intonation, which are sometimes based on auditory impressions or limited empirical research to support phonological claims. Contour clustering is applicable to spontaneous and scripted speech of any language. There is no restriction as to which prosodic domain (intonation unit, (intermediate) phrase, word, syllable) can be investigated and there is limited need for annotation prior to analysis. The core of this approach is a cluster analysis on time-series of f0 measurements and consists of two scripts (Praat and R). Graphical user interfaces can be used to perform the analyses and speaker variability can be accounted for. As determining the number of clusters is a key part of the analysis, graphical feedback (plots) is provided for each clustering round (example in Figure 1). After cluster analysis, Praat textgrids can be generated with the cluster number annotated for each individual contour in the data. Although further confirmatory analysis is still required, the outcomes provide useful and unbiased directions for any investigation of prototypical f0 contours based on their acoustic form. These features make the approach particularly useful for language documentation, where the description of prosody is often lacking.

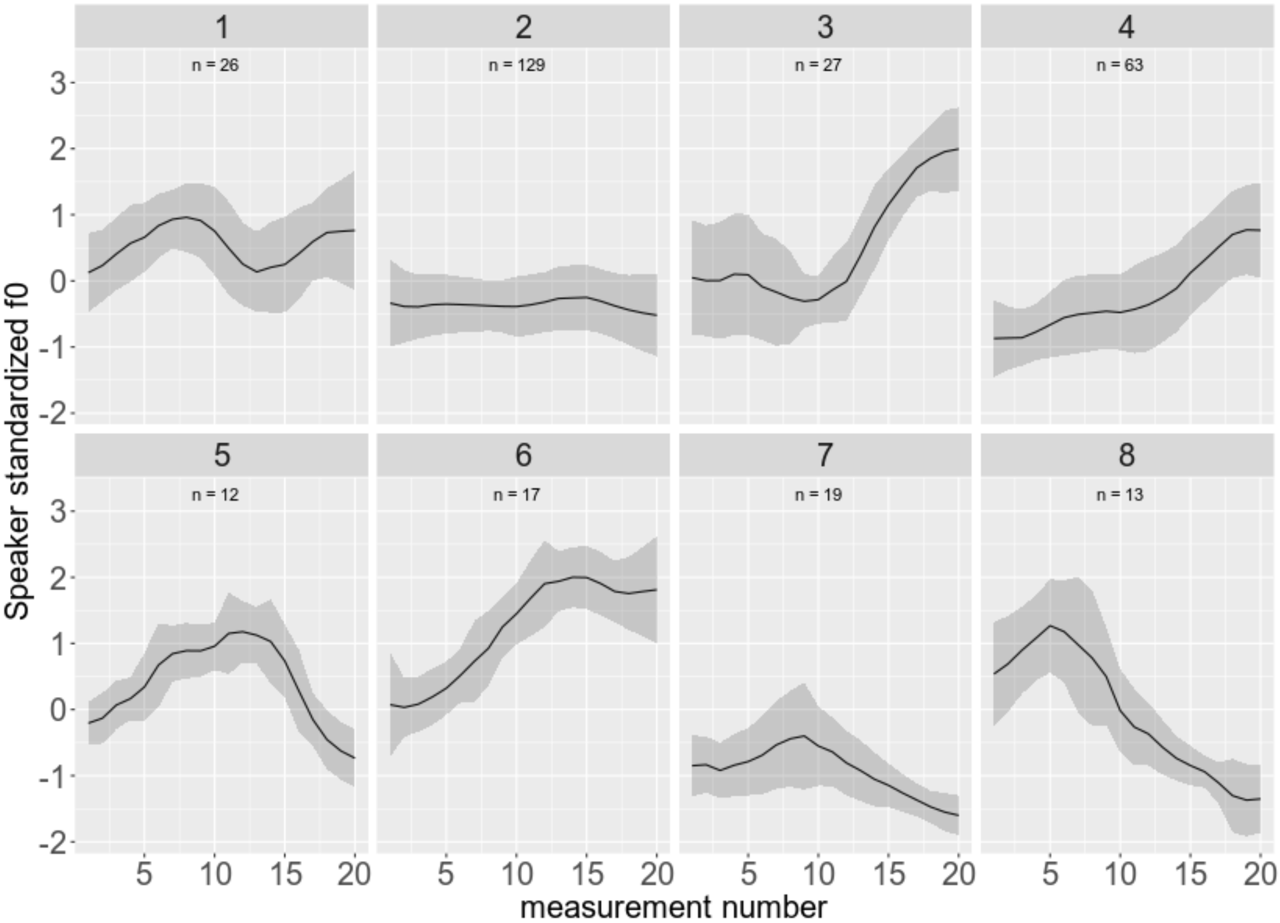

Figure 1. Example of a graphical output based on spontaneous Papuan Malay data assuming eight clusters. Each contour represents the mean speaker standardized f0 (and standard deviation) in a cluster.

Automatic Analysis of Speech Prosody in Dutch

Na Hu[1], *Berit Janssen*[2], *Judith Hanssen*[3], *Carlos Gussenhoven*[4], *Aoju Chen*[1]

[1] Utrecht University, the Netherlands
[2] Digital Humanities Lab, Utrecht University, the Netherlands
[3]Avans University of Applied Sciences, the Netherlands
[4] Radboud University, the Netherlands
{n.hu, b.d.janssen, aoju.chen}@uu.nl, judithhanssen@gmail.com, c.gussenhoven@let.ru.nl

In this talk we present the first publicly available tool for automatic analysis of speech prosody (AASP) in Dutch. Incorporating the state-of-the-art analytical frameworks, AASP enables users to analyze prosody from two different theoretical perspectives. Structurally, AASP analyzes prosody in terms of prosodic events within the auto-segmental metrical framework, hypothesizing prosodic labels in accordance with Transcription of Dutch Intonation (ToDI). Holistically, by means of the Functional Principal Component Analysis (FPCA) AASP generates mathematical functions that capture changes in the shape of a pitch contour. Regarding ToDI, AASP performs four tasks including pitch accent detection, pitch accent classification, prosodic boundary detection, and prosodic boundary tone classification. Using SVM, AASP performs with accuracy comparable to similar tools for other languages for pitch accent detection, prosodic boundary detection, and prosodic boundary tone classification. Notably, we have found that by combining functional features extracted from FPCA with conventional acoustic features, AASP can attain a higher accuracy for pitch accent classification (76.87%) than AuToBI for English using conventional acoustic features (71.6%). Regarding FPCA, AASP outputs the weights of principal components that capture core variations in the shape of pitch contours in a .csv file, which can be directly used for further statistical analysis.

Published as a Docker container, AASP can be set up on various operating systems in only two steps. Moreover, the tool is accessed through a graphic user interface, making it accessible to users with limited programming skills. It has also the potential to be adapted for prosodic analysis in other languages.

Listeners learn and predict talker-specific prosodic cues in speech perception

Giulio G.A. Severijnen [1], Hans Rutger Bosker[2,] Vitória Piai[1, 3], & James M. McQueen[1,2]

[1] Donders Centre for Cognition

[2] Max Planck Institute for Psycholinguistics

[3] Donders Centre for Medical Neuroscience

One of the challenges in speech perception is that listeners must deal with considerable segmental and suprasegmental variability in the acoustic signal due to differences between talkers. Most previous studies have focused on how listeners deal with *segmental* variability. In this EEG experiment, we investigated how listeners track talker-specific usage of *suprasegmental* cues to lexical stress to correctly recognize spoken words. In a 3-day training phase, Dutch participants learned to map non-word minimal stress pairs onto different object referents (e.g., *USklot* means “lamp”; *usKLOT* means “train”*)*. These non-words were produced by two male talkers. Critically, each talker only used one suprasegmental cue to signal lexical stress (e.g., Talker A only used F0, Talker B only amplitude). We expected participants to learn which talker used which cue to signal stress. In the test phase, participants indicated whether spoken sentences including these non-words were correct (“The word for ‘lamp’ is...”). We recorded participants’ response times and EEG patterns, targeting an ERP related to phonological prediction: the N200. We found that participants were slower to indicate that a stimulus was correct if the non-word was produced with the unexpected cue (e.g., Talker A using amplitude). That is, if in training Talker A used F0 to signal stress, participants experienced a mismatch between predicted and perceived phonological word-forms if, at test, Talker A unexpectedly used amplitude as cue to stress. This illustrates talker-specific prediction of suprasegmental cues, picked up through perceptual learning in training. In contrast the N200 amplitude, was not modulated by the mismatch. Theoretical implications for these results are discussed.

Prosodic phrasing of short left-dislocated adverbial adjuncts in Brazilian Portuguese

Tainan Carvalho[1], Luciani Tenani[1], and Marc Swerts[2]

We explore the prosodic configuration of short left-dislocated adverbial adjuncts in Brazilian Portuguese (BP), as "*Amanhã*" in "*Amanhã*, nosso juiz decidirá o caso" – for the English "*Tomorrow* our judge will adjudicate the case". Specifically, we discuss how the prosodic configuration of these constituents changes depending on whether they represent neutral or topicalized adverbial adjuncts. Our hypothesis is that the short left-dislocated adverbial adjuncts induces an intonational phrase (IP) boundary when it is topicalized, but not when it occurs after a neutral adjunct. We analyzed speech recordings from thirteen BP speakers (all female, native speakers of São Paulo State variety of BP), and measured a set of phonetic cues that have previously been associated with IP boundaries: pause, duration and F0 variation. Our speakers were asked to read (three times) a set of utterances with left-dislocated adverbial adjuncts. Adverbial adjuncts utterances were included in broader contexts that were semantically manipulated to favor both neutral and topicalized readings. The contexts were randomized and mixed with distractors. The results confirm that the prosodic phrasings of the short adverbial adjuncts depended on the context. Topicalized dislocated adverbs appear to differ from neutral ones in that: (i) they are more frequently marked by the occurrence of final boundary tones (H% and L%); (ii) the pauses post adjunct are more frequent in topicalized contexts, and also appear to be longer on average than in neutral contexts (topicalized: 95 ms; neutral ones: 39ms) and (iii) preboundary lengthening occurs more consistently in topicalized adverbial adjuncts.

[1] São Paulo State University, UNESP
[2] Tilburg University

And now for something completely different…

Vincent J. van Heuven
Pannon Egyetem. Veszprém
v.j.j.p.van.heuven@hum.leidenuniv.nl

I will discuss a recent court case in the Netherlands, in which forensic phonetic expertise was called upon to help settle a dispute over trade name infringement. In 2014, Dutch brewer Grolsch launched a beer called *Kornuit* /kɔrˈnœyt/. Recently, supermarket chain Lidle released a beer under the name *Kordaat* /kɔrˈdaːt/. I was asked by Grolsch to shed light on the phonetic similarity between the brand names. Using the Levenshtein distance metric (Levenshtein 1966, Heeringa 2004), the phonetic difference between the names is 29 percent. To show that the similarity between the brand names was very likely to be intentional rather than accidental (as Lidle would have it), I established the statistical distribution of the similarity of Dutch word pairs. I selected the 3000 most frequent monomorphemic content words from Baayen et al. (1995) and computed the Levenshtein distance for all 4,498,500 non-identical word pairs (using Gabmap software, Leinonen et al. 2016). Distances ≤ 29% occur in .5 percent of the word pairs, which arguably shows that the name *Kordaat* was not accidentally chosen by Lidl. In my talk I will explain the Levenshtein metric and motivate the decisions made to obtain the distribution of distances between Dutch word pairs.

References

Baayen, R. H., Piepenbrock, R. & Gulikers, L. (1995). CELEX2 LDC96L14. Web Download. Philadelphia: Linguistic Data Consortium.

Heeringa, W. J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Doctoral dissertation, University of Groningen.

Leinonen, T., Çöltekin, Ç. & Nerbonne, J. (2016). Using Gabmap. *Lingua*, 178, 71-83.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10*(8), 707-710.