Listener-specific perception of speaker-specific productions in intonation

Francesco Cangemi, Martina Krüger, Martine Grice
Universität zu Köln, I*f*L Phonetik

We see earth by earth, water by water
Bright aether by aether, and obliterating fire by fire
Love by love, and strife by baneful strife
Empedocles, Fragment 109

Abstract

In this contribution we explore the hypothesis of an interaction between speaker- and listener-specific strategies in the encoding and decoding of intonational contrasts. Intonational categories, such as the pitch accents used in the signalling of focus types, can be cued by different phonetic exponents, such as peak alignment or duration of the target words. Through a production task we document speaker-specific strategies: Individual speakers might use more or fewer cues than others (robustness) when encoding intonational contrasts, and each cue can be used to encode one or more contrasts (partitioning). We show in a subsequent perception task that listeners are sensitive to speaker-specific strategies, since correct identification scores for productions of individual speakers mirror the robustness and partitioning of speakers' productions. Moreover, listeners vary as to how reliably they decode intonational contrasts across speakers. However, in line with the hypothesis of an interaction between speaker- and listener-specific behaviours, some listeners are more reliable at decoding contrasts as encoded by some particular speakers, which in turn are decoded less reliably by other listeners. These findings suggest that phonetic cues to intonational contrasts should not be understood as singly necessary and jointly sufficient features for category membership, but rather as dimensions along which phonological categories cluster, in an individual-specific network of phonological knowledge.

1. Introduction

1.1. Background

It is not an overstatement to say that in recent years phonetic research has stepped away from the brutal averaging of data points collected across subjects, be it in perception or production, although this has only just begun to play a role in research into intonation. We begin by taking the example of the voicing contrast in plosives, particularly in syllable initial position. In their seminal paper, Lisker and Abramson (1964) collected acoustic datafrom 11 languages, represented by only 17 speakers altogether. Despite the fact that their four speakers of American English showed massive differences in their individual behaviour (Lisker and Abramson 1964: 538), voicing in Dutch, Tamil, Cantonese, Eastern Armenian, Korean, Hindi and Marathi was investigated using data from a single speaker for each language. Later on, in his groundbreaking study on the phonetic exponents of voicing in word-internal stops, Lisker (1986) reviewed 16 acoustic cues to voicing, again under the tacit assumption that the weighting of such cues would not be affected by listener-specific patterns. Individual specificity in production and perception was thus out of the picture in these two studies, which focussed on cross-language comparisons and on the relationships between articulation and perception, respectively.

In recent years, however, speaker- and listener-specific behaviour has gained a central role in the study of how phonetic substance maps onto phonological contrasts. This evolution might have stemmed from the ability of linguists to integrate insights and practices from neighbouring disciplines, both at the theoretical level (as in the case of a renewed understanding of category structure, e.g. Lakoff 1987) and at the methodological level (as with mixed-effects modelling, notably through the targeted exploration of random coefficients, e.g. Baayen 2008). As a consequence, recent studies on voicing contrasts in stops have devoted a great deal of attention to speaker- and listener-specific behaviour - not only as important factors in the data analysis, but also as dimensions shaping the actual research questions. Allen et al. (2003), for example, document systematic variation of voice onset time patterns in stop contrasts across speakers, and link this finding to speaker recognition mechanisms. Individual differences are found in the weighting of the cues associated with stop contrasts in production (e.g. Schultz et al. 2012, for voicing in English) and perception (e.g. Idemaru et al. 2012, for stop length in Japanese). Research on individual behaviour has also been conducted in the effort to provide evidence in favour of theories suggesting a link between production and perception. Recent studies in this vein include Perkell et al. (2004a, b), which tested the hypothesis that the more precisely a subject discriminates a contrast as a listener, the more accurately that subject will produce such contrast as a speaker, both in terms of articulation patterns and acoustic output. Findings from Bradlow et al. (1996), Newman et al. (2001), Hazan and Baker (2011) and Hazan et al. (2013) are also compatible with the assumption of more accurate production resulting in greater intelligibility. Speaker and listener-specific behaviours are thus well attested for segmental contrasts, which have been studied extensively in the past fifty years.

The situation for intonation (and prosody in general) is radically different. Despite the fact that there is an abundance of studies reporting on language-specific marking of focus types using accent types, deaccentuation or dephrasing (e.g. Jun 2014), only few studies focussed on individual-specific differences, notably in production. An unpublished study by Andreeva and Barry (2007) on phrasal prominence suggests that its realization differs not only across the investigated languages (Bulgarian and Russian), but also between the speakers of each of the two languages. Niebuhr et al. (2011) show that an intonational contrast in Standard Northern German is cued by one group of speakers through differences in peak alignment, and by another group through differences in peak shape. This paper did not investigate the consequences of these different production strategies for perception. In fact, to our knowledge, no studies have targeted listener-specific strategies in the decoding of intonational contrasts – let alone the interaction between specificity in production and perception.

1.2. Rationale

In this contribution, we explore the *interaction* between speaker- and listener-specific behaviour in the encoding and decoding of prosodic categories. Note that this is different from exploring the *link* between speaker- and listener-specific behaviour, as in the studies by Perkell et al. (2004a,b) cited above, in which subjects participated in both a production and a perception task. Their results showed that some individuals produce contrasts more accurately than others, that some individuals discriminate contrasts more pecisely than others, and that accurate speakers are also precise listeners (see Fig. 1).

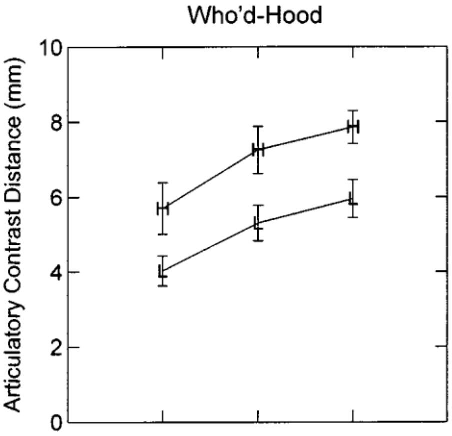
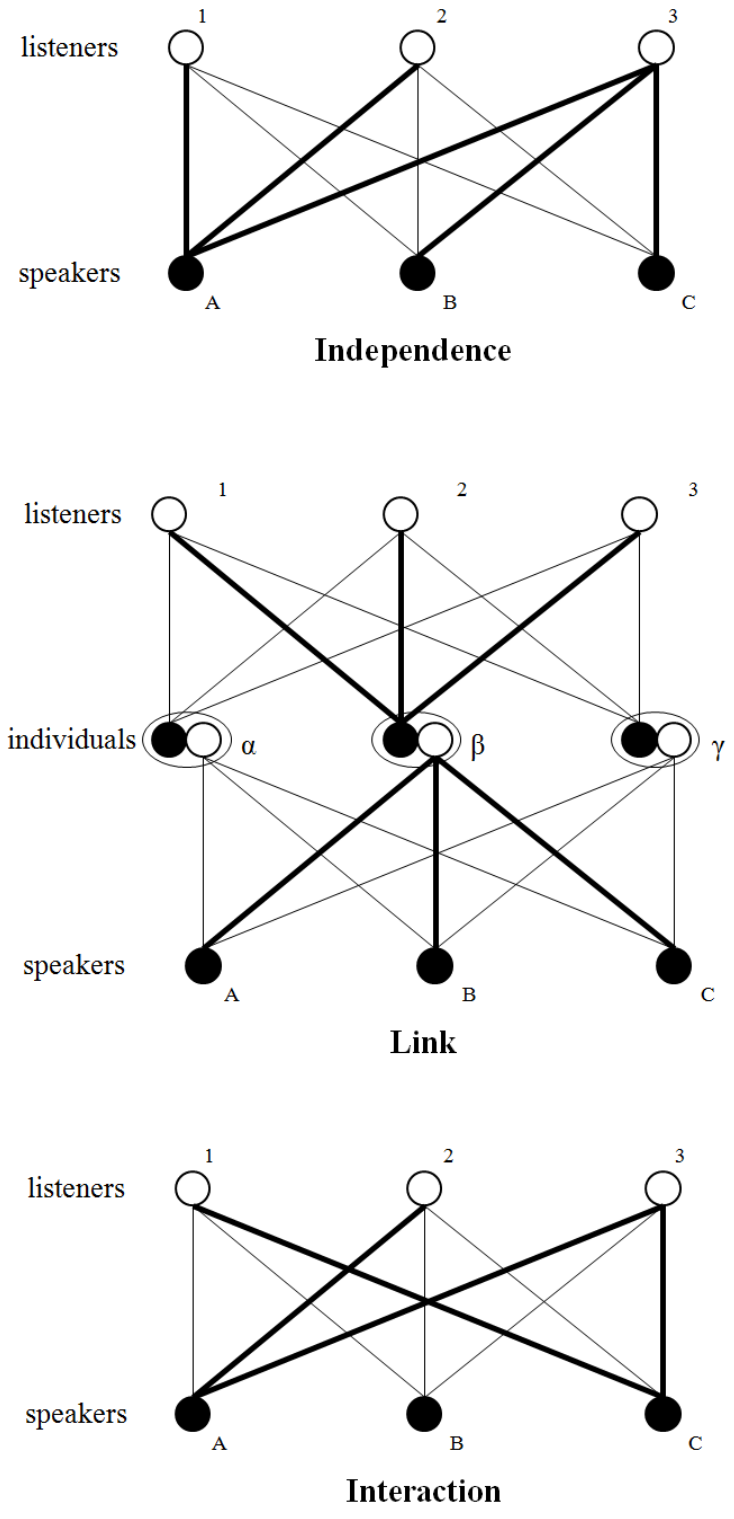


Fig. 1: Articulatory contrast distance for tongue body position (y-axis; error bars are one standard error about the mean) as a function of three speaking conditions (x-axis; Fast, Normal, Clear) for

the /ʊ,u/ contrast. Subjects are split into two groups based on their performance in a two-step discrimination task involving stimuli on the [ʊ,u] continuum. Listeners are in the “high discrimination ability” (H) group if their responses are 100% correct, otherwise they are in the “low discrimination ability” (L) group (readapted from Perkell et al. 2004a: 2343, Fig. 4).

This methodology is particularly suited for documenting how good individuals are at producing and perceiving contrasts – that is, at profiling the “best speakers” and “best listeners”, somehow assuming that there is a phonetic equivalent of the blood-type notions of “universal donors” and “universal receivers”. Our aim is to show that not only some speakers might be generally more accurate and thus more intelligible than others, but also that some speakers might produce contrasts in a way that make them easily intelligible *to some particular listeners*, but not to others. This would document an interaction, rather than a link, between specificity in production and perception.

Figure 2 sums up the three potential scenarios. The behaviour of speakers and listeners could be independent, it could be linked or it could interact. In each panel, individuals are represented by nodes. Listeners are represented by empty circles identified by numbers and speakers are represented by filled circles identified by letters. The association lines between nodes represent how reliably a contrast produced by a given speaker is perceived by a given listener. Thick lines indicate that the intended categories produced by the speakers are frequently perceived correctly by the listener; thin lines indicate that this is rarely the case.



Independence

Link

Interaction

Fig. 2: Independent, linked and interacting speaker- and listener-specific behaviour.

The top panel illustrates a situation in which some speakers (filled circles) are overall more intelligible and some listeners (empty circles) are overall more reliable, as indicated by the number of thick lines departing from each node. Crucially, the two phenomena are *independent*. Speaker A's productions are reliably perceived by all listeners, whereas listener 3 reliably perceives productions from all speakers. Using the blood types metaphor introduced above, speaker A would thus be an example of a "universal donor" and listener 3 would be an "universal recipient".

The mid panel depicts a *link* between production and perception within individuals. This is akin to the results reported by Perkell et al. (2004a,b), in which subjects participated in both a production and a perception task. The mid panel thus features three tiers, since subjects (in the central tier) serve as both listeners and speakers (hence the juxtaposed filled and empty circles identified by Greek letters), and thus both listeners (top tier) and speakers (bottom tier) are required. The illustration shows that some individuals (i.e. node β) are both accurate in their productions (thick lines connecting to listeners in the top tier) and reliable in their perceptual judgements (thick lines connecting to speakers in the bottom tier). Using the blood types metaphor introduced above, the individual β would thus be an example of an individual who is at the same time a "universal donor" and a "universal recipient".

The bottom panel illustrates the presence of an *interaction* between speaker- and listener-specific behaviour. Some speakers might still be overall more intelligible than others, and the same might apply for listeners. Crucially, however, there is no such thing as a "universally intelligible speaker" or a "universally proficient listener", as in the *independence* scenario, and thus (a fortiori) no individual who is both, as in the *link* scenario. Rather, a listener might perceive more reliably the contrasts produced by a given speaker, whose productions are in turn perceived less reliably by a different listener. Similarly, a given listener might be very reliable at decoding productions from a particular speaker, but perform very badly on productions from a different speaker. This is exemplified by speaker A being badly perceived by listener 1, who however is very reliable at decoding contrasts produced by speaker C. This is still compatible with some individual being overall better listeners (e.g. listener 3, with two thick lines departing from its node) or worse speakers (e.g. speaker B, with no thick lines departing from its node), as in the *independence* and *link* scenarios. However, in the *interaction* scenario these main effects can be modulated by specific interactions, and thus neither accuracy in production nor precision in perception need to be understood in absolute terms.

In the following sections, we explore the hypothesis of an interaction between speaker- and listener-specific strategies, using a dataset on the production and perception of focus in German collected for various purposes (Mücke and Grice 2014, focussing on production; Grice et al. in prep., focussing on perception). Before providing an analysis of the interaction between specific speaker and listener behaviours (3.3), we thus summarise some of the relevant aspects of the two original studies.

2. Methods

2.1 Production task

Participants and recordings. Recordings were made of five speakers (three female) of Standard German from north of the Benrather isogloss, aged between 22 and 37 years. Articulatory movements were captured with a 2D Electromagnetic Articulograph (Carstens AG 100), with sensors on the upper and lower lips, recorded at 500Hz, downsampled to 200Hz and smoothed with a 40Hz low-pass filter. Simultaneous acoustic recordings were made with a DAT-recorder (TASCAM DA-P1) using a condenser microphone (AKG C420 head set) and sampled at 44.1kHz, 16bit.

Materials. The materials contained target words /ˈbiːbɐ/, /ˈbaːbɐ/, and /ˈboːbɐ/ (fictitious names). These names were in the default position for the nuclear pitch accent (the last argument of the verb). Information structure was manipulated by means of question-answer pairs. Four different focus structures were elicited: the target word occurred either as part of the background or in broad, narrow or contrastive focus. An example of a set of question-answer pairs is given in Figure 3 for the target word <Bahber> /ˈbaːbɐ/.

Questions:

1. Will Norbert Dr. Bahber treffen? *Does Norbert want to meet Dr. Bahber?*
2. Was gibt's Neues? *What's new?*
3. Wen will Melanie treffen? *Whom does Melanie want to meet?*
4. Will Melanie Dr. Werner treffen? *Does Melanie want to meet Dr. Werner?*

Answers:	**test word in:**
Melanie will Dr. Bahber treffen.	
1. [———]$_{focus}$	background
2. [————————————————]$_{focus}$	broad focus
3. [————]$_{focus}$	narrow focus
4. [————]$_{focus}$	contrastive focus

(lit.: *Melanie wants Dr. Bahber to-meet*)

Fig. 3: Speech material example, target word <Bahber> /ˈbaːbɐ/, taken from Mücke & Grice 2014: 52.

Subjects were presented with the contextualizing question both auditorily and visually. They then read aloud the answer in a contextually appropriate manner at a speaking rate which they considered to be normal. Question-answer pairs were randomized to avoid repetitions in sequences. In total, 560 tokens were recorded (4 target words x 4 focus structures x 7 repetitions x 5 speakers), although only 420 tokens were analysed (one target word having been discarded, owing to difficulties identifying lip aperture in the articulatory analysis).

Labels and measurements. Intonation was transcribed by two annotators using the acoustic waveform and F0 contours in PRAAT (Boersma & Weenink, 2010). In cases where transcribers differed (16%), a consensus transcription was reached. Accented target words were labelled using one of three different GToBI accent types (Grice et al. 2005): H+!H*, H* and L+H*, as presented schematically in Figure 4a-c. In all cases there was a low boundary tone sequence following, labelled as L-% (equivalent to L-L% in ToBI for English).

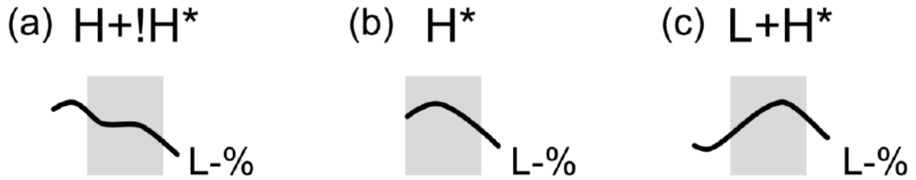


Fig. 4: Schematic representation of the three different pitch accent types as presented in the GToBI online guidelines http://www.gtobi.uni-koeln.de.

Acoustic durations (target words and stressed syllables) were labelled by hand using the EMU speech database system (Cassidy & Harrington 2001). For the kinematic recordings, the lip aperture index (LA, Byrd 2000: 6) was calculated in terms of the Euclidean distance between the two sensors on the upper and lower lips capturing movements both in the horizontal and vertical dimensions. Kinematic labels were identified corresponding to the lip opening gesture in the stressed syllable, i.e. the movement from the maximum lip closure in the onset consonant to the maximum opening of the lips in the vowel, and the point of maximum velocity.

2.2 Perception task

Participants. Twenty native speakers of German (20 to 40 years of age, mean 25.9 years) with no knowledge of linguistics participated in the perception experiment. Participants had self-reported normal hearing.

Materials. Test sentences were taken from the production study described above. Thus, the carrier sentence “Melanie will Doktor _____ treffen.” (Melanie wants to meet Doctor _____) contained one of the three fictitious target names: Bieber, Bahber and Bohber from one of the five speakers in one of the four focus conditions. Experimental materials contained the 420 tokens from the production study (3 target words x 5 speakers x 4 focus structures x 7 repetitions) plus 60 stimuli from a practice phase (4 focus structures x 3 target words x 5 speakers).

Every speaker was evaluated in a separate block, and within each speaker-block, every target name (Bieber, Bahber, Bohber) was evaluated separately. Target word blocks were randomized within the speaker blocks, and the speaker blocks were also randomized for each participant (controlled permutation). This allowed a controlled order of speaker blocks that was counterbalanced in order to avoid a possible influence on the judgments.

A practice phase of twelve stimuli (4 focus structures x 3 target words) preceded each speaker-block in order for the participants to familiarize themselves with the procedure and with possible speaker-specific strategies. For this practice phase, 12 prototypical stimuli were selected for each speaker, each target word and each focus structure. These items were those consistently assigned to the correct question/focus structure by six trained phoneticians in a pretest. In order to minimize learning effects, a given target word was only included in a single focus condition for each speaker. Practice phase stimuli were excluded from further analysis.

Procedure. The experiment was conducted with the PARADIGM software (Perception Research Systems 2007). Instructions were given in written form. The task was to match the test sentences heard to one of four questions (see Figure 3) presented on the screen. This was done by clicking on the question that subjects judged to be the most appropriate for a particular test sentence. There was no time limit for the choice.

In order to assure the comprehension of the task, participants were asked in a pretest to produce the target sentence (Melanie will Dr. Bahber treffen) as an answer to the four questions asked by the experimenter. None of the subjects reported difficulties in carrying out the task. Participants heard every test sentence once via headphones. The test sentences were preceded by a beep in order to assure full attention.

3. Results

3.1. Production

Table 1 shows a synopsis of the acoustic analysis, split by cues (rows) and speakers (columns); results refer to the three focus types (Broad, Narrow and Contrastive). Each cell shows how a given speaker uses a given cue in the encoding of the focus types; the tilde indicates absence of statistically significant differences between focus types[1]. Cells are displayed in different shades of grey according to the number of contrasts between focus types that a given cue allows for a given speaker. For example, peak height is significantly different for the three focus conditions in productions from speaker F3 (dark grey); for speaker M1, peak height is only significantly different in Broad focus cases, compared to both Narrow and Contrastive focus (light grey); for speaker M2, peak height does not vary across the three focus conditions (white). Speaker-specific differences are evident in both terms of *robustness*, involving the number of cues used in the encoding of the three categories, and in terms of *partitioning*, that is, whether a given cue is used to distinguish between two or more categories.

[1] Significance at p = 0.05 was assessed through ANOVAs run separately for each speaker and cue, and followed by post-hoc Tukey’s HSD tests. For details on the quantitative analysis, including the direction of the reported effects and results for the background condition, see Grice et al. (in prep.).

In Table 1, the number of white cells for each speaker gives a measure of robustness, in terms of how many cues are used to encode focus contrasts. Whereas speakers F1, F2 and M1 use all five explored cues, speaker F3 only uses three, and speaker M2 only uses two (i.e. duration of target word and number of prenuclear accents). The number of dark grey cells for each speaker can be seen as a measure of partitioning – that is, whether the phonetic space of the cue is partitioned into multiple regions (each corresponding to a category). For example, the duration of the target word is significantly different in the three focus conditions for speakers F3 and M1, but only allows for a single contrast in productions from speakers F1 and F2 (differentiating cases of Contrastive focus from cases of Broad or Narrow focus) and from speaker M2 (for whom duration rather differentiates cases of Broad focus from cases of Narrow or Contrastive focus).

Interestingly, Table 1 shows that the contrast between focus types is encoded by all speakers, albeit with different degrees of robustness and partitioning. In productions from speaker M2, for example, Broad focus can be distinguished from Narrow and Contrastive focus through a single cue (viz. the acoustic duration of the target word), and Contrastive focus can be distinguished from Broad and Narrow focus through a single other cue (viz. the number of prenuclear accents). While this means that the three intended categories can still be reliably decoded through their acoustic exponents, it is clear that this speaker encodes the three-way contrast in a suboptimal way – especially if compared with productions from speaker F2, who differentially encodes categories using all cues (high robustness), two of which (peak alignment and height) actually allow for three-way contrasts (high partitioning).

Speaker **Cue**	F1	F2	F3	M1	M2
Peak alignment	B N~C	B N C	B N C	B N~C	B~N~C
Peak height	B N~C	B N C	B N C	B N~C	B~N~C
Duration of target word	B~N C	B~N C	B N C	B N C	B N~C
Duration of first word	B N~C	B N~C	B~N~C	B N~C	B~N~C
Number of prenuclear accents	B N~C	B~N C	B~N~C	B N~C	B~N C

Tab. 1: Encoding of contrasts between three focus categories (Broad, Narrow and Contrastive), split by cues (rows) and speakers (columns). The tilde indicates absence of statistically significant differences between focus categories.

The articulatory analysis provides comparable results. Figure 5 shows averaged lip aperture trajectories broken by speaker (columns), target words (rows) and focus conditions (line types). Again, trajectories are clearly distinguishable for speaker F1, for all four focus conditions, in all target words. This is not the case for productions from speaker F3, for whom only one out of four focus conditions (viz. contrastive) seems to follow a different pattern, and only for two out of three target words (Bahber and Bieber).

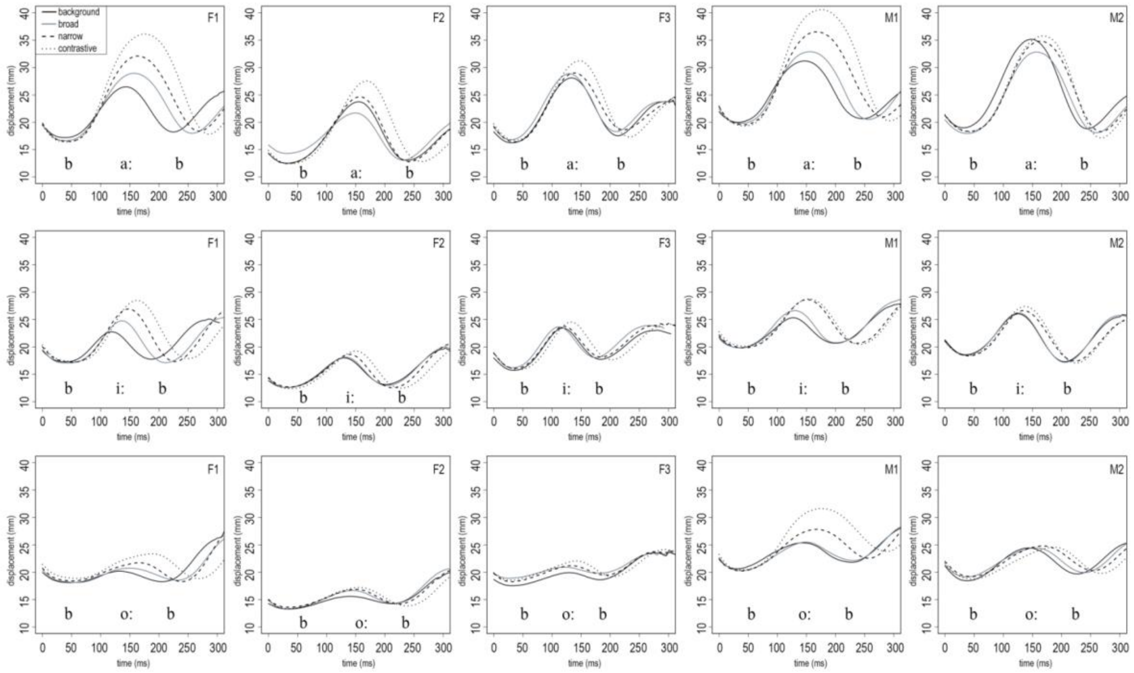


Fig. 5: Averaged trajectories for the target words B/aː/ber, B/iː/ber and B/oː/ber, separately for each speaker (F1, F2, F3, M1, M2) with different focus structures. All trajectories are aligned with the acoustic beginning of the target word. (Fig. 10 from Mücke & Grice 2014: 56, original figure and caption)

3.2. Perception

Results from the perception task are presented as percentages of listeners' correct responses, evaluated with respect to the intended categories produced by the five speakers above (chance level: 25%). Figure 6 shows responses pooled across listeners and split by speakers. The trends are consistent with the expectations stemming from the production study. For instance, productions from speaker F2, who encoded focus robustly and distinctively, are correctly identified more often (69.32%) than productions from speaker M2 (63.7%), which had suboptimal encoding of focus (cf. 3.1). Speakers can thus be arranged along a continuum of contrast maximization in encoding focus structures. This result is not incompatible with the notion of a "universal phonetic donor", that is, of a speaker being generally more accurate in encoding phonological contrasts, which in turn makes such contrasts easier to decode for all listeners (cf. 1.2).

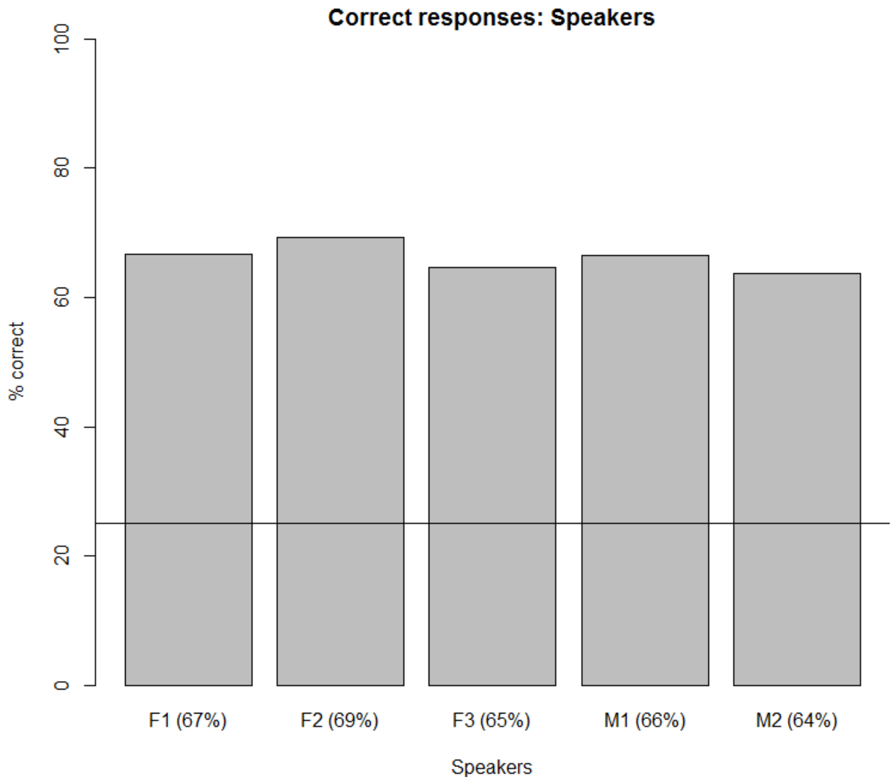


Fig. 6: Percentage of correct responses from all listeners to stimuli produced by individual speakers. The horizontal line indicates chance level.

When split by individual listeners (Figure 7), results from responses to productions from all speakers indicate an even greater variability in how proficient individuals are at decoding focus structures, with one listener providing correct answers in three out of four cases (viz. BB, 74.87% correct) and another listener in just over half of the cases (viz. KS2, 55.55% correct). This is, again, compatible with the notion of a "universal phonetic recipient", that is a listener who is overall more reliable at decoding intended categories as produced by any speaker.

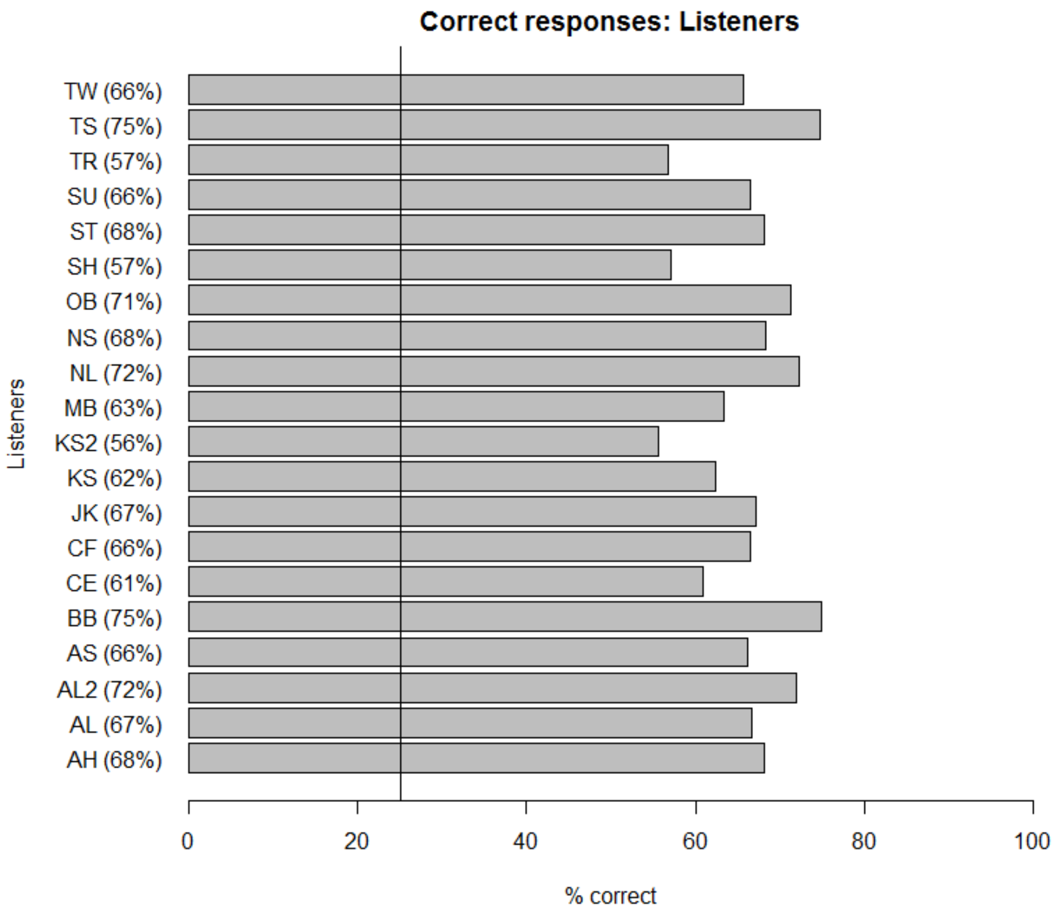


Fig. 7: Percentage of correct responses to stimuli from all speakers for individual listeners. The vertical line indicates chance level.

3.3. Interaction

The qualitative analysis of the results from the perception study is thus compatible with the notion that some speakers (listeners) are overall more proficient in encoding (decoding) focus structures. In the following, we provide a quantitative evaluation of the hypothesis that some particular listeners might be particularly proficient at decoding structures as encoded by some particular speakers - that is, the hypothesis of an interaction between speaker- and listener-specific behaviour (cf. Figure 2, bottom panel).

The heat map in Figure 8 shows correct responses pooled across focus conditions and split by speakers (x-axis) and listeners (y-axis), with darker shades of grey corresponding to higher correct response scores. Average scores pooled across listeners and speakers are put in parentheses after each speaker and listener identifier on the axes, thus incorporating information from Figures 6 and 7. If any single speaker had been more intelligible to all listeners overall, we would expect one single column in the figure to be darker than the others. Similarly, had any single listener been more successful at decoding contrasts produced from all speakers, we would expect the presence of continuous darker rows in the figure. An informal analysis of the figure, however, shows that this is not the case. It is true that some columns might seem overall darker than others, thus indicating that a given speaker is more intelligible than another (e.g. F2 vs. M2), as confirmed by the average scores on the x-axis and by Figure 6. It is also true that some rows seem overall darker than others, thus indicating that a given listener is more successful than another (e.g. BB or TS vs. KS2 or SH or TR), as confirmed by the average scores on the y-axis and by Figure 7.

But Figure 8 also shows a more interesting pattern of results: The same speaker can produce contrasts which are well decoded by a certain listener, but poorly decoded by another listener.

Productions from speaker F1, for example, are decoded very reliably by listeners BB and ST, but very poorly by listeners SH and CE. Similarly, the same listener can reliably decode contrasts as produced by a given speaker, while being less reliable with productions from a different speaker. Listener MB is for example very reliable when decoding contrasts produced by speaker F2 but is less reliable with productions from speaker F1.

Crucially, the same speakers and listeners can be involved in diametrically opposed patterns of results: whereas listeners CE and AL seem to over-perform on productions from speakers M1 and underperform on productions from speaker F1, listeners ST and JK seem to do the opposite (over-performing on F1 and under-performing on M1). An informal analysis of Figure 8 is thus consistent with the hypothesis of an interaction between speaker- and listener-specific behaviour.

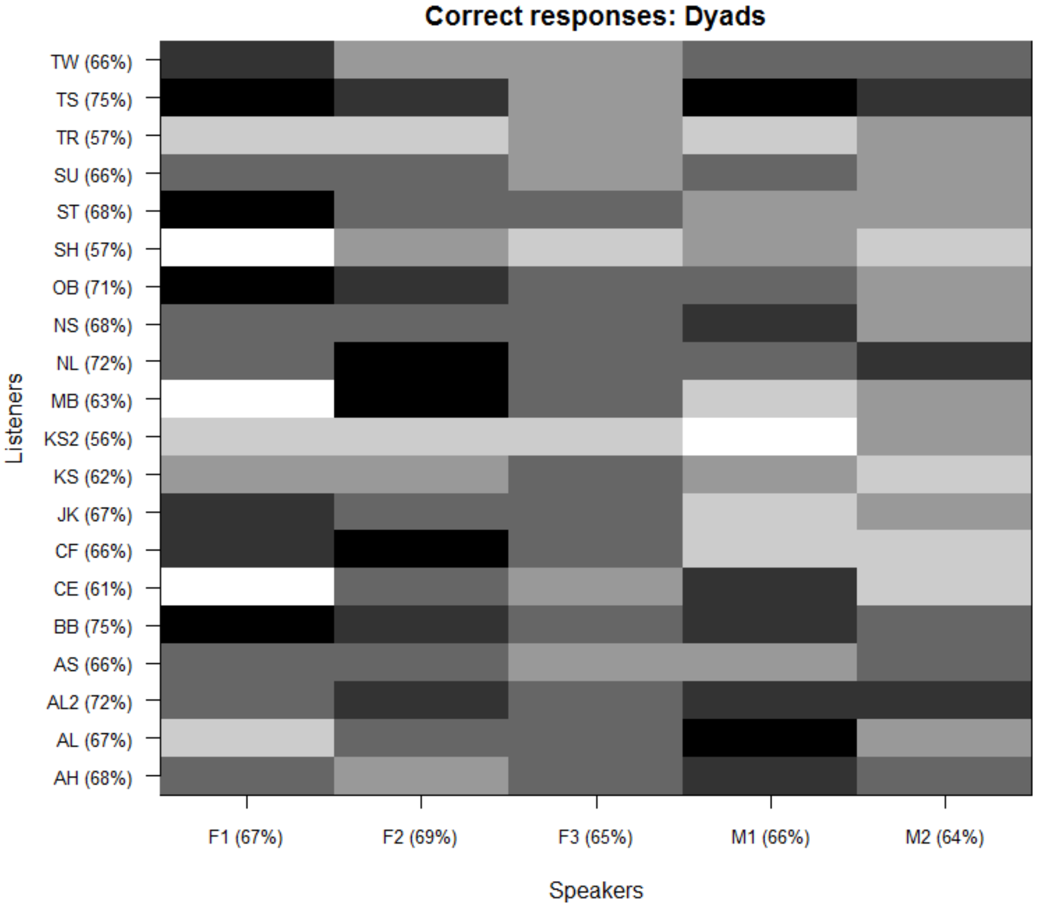


Fig. 8: Percentage of correct responses for individual speakers from individual listeners.

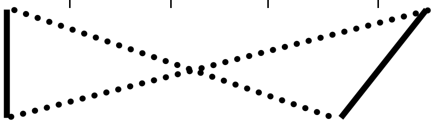
The most straightforward way to test this hypothesis is to conceptualize such interaction as an interaction in the statistical sense. We thus used logit modelling to predict correct identification scores (on data from all focus categories) using the factors SPEAKER (from 1 to 5), LISTENER (from 1 to 20), and their interaction. This full model was compared to a null model which dropped the interaction between the two factors. A Likelihood Ratio Test yielded highly significant results ($\chi^2(76)=145.46$, p=0.000003).

Significant results were achieved also through testing based on mixed effect models. The null model included random intercepts for SPEAKER (from 1 to 5) and LISTENER (from 1 to 20) only. The full model also included random intercepts for DYAD (from 1 to 100), that is the individual pairings of speakers and listeners (e.g. speaker F1 with listener AH, speaker F1 with listener AL, et cetera). A Likelihood Ratio Test revealed a significant difference between the two models ($\chi^2(1)=18.884$, p=0.00002).

In order to quantify the dyadic interaction effects illustrated through a grey scale in Figure 8, we ranked the 100 random intercepts for DYAD, assigning the first place to the most beneficial interaction (which indicates that the listener in that dyad is particularly proficient at decoding contrasts as encoded by the speaker in that dyad) and the last place to the most detrimental. We conservatively focussed on the 10 most detrimental interactions (with DYAD random intercepts below -0.2, rankings from 91 to 100) and on the 10 most beneficial interactions (with DYAD random intercepts above 0.2, rankings from 10 to 1) only. Table 2 shows rankings, coefficients and dyads (relevant identifiers with speakers in boldface and listeners in italics, separated by a colon).

most detrimental ←

Rank	100	99	98	97	96	95	94	93	92	91	90	89
Coef.	-0.39	-0.33	-0.28	-0.22	-0.21	-0.21	-0.21	-0.2	-0.17	-0.17	...	...
Dyad	F1:CE	F1:MB	**F1**:*AL*	M1:CF	F1:SH	M1:KS2	**M1**:*JK*	M2:CF	M2:OB	F3:SH	...	...



Dyad	...	...	**F1**:*JK*	M1:CE	F1:TS	**M1**:*AL*	F2:CF	F1:BB	F1:OB	F2:NL	F2:MB	F1:ST
Coef.	...	...	0.2	0.23	0.25	0.25	0.26	0.28	0.3	0.3	0.3	0.31
Rank	12	11	10	9	8	7	6	5	4	3	2	1

→ most beneficial

Tab. 2: Random intercepts for Dyads.

Interestingly, we found for example that listener AL is remarkably reliable at decoding productions of speaker M1 (ranking at 7) but also unreliable at decoding productions by speaker F1 (rank 98), whereas listener JK has the opposite behaviour, being particularly reliable with productions from speaker F1 (rank 10) and unreliable with M1 (rank 94). The solid lines in Figure 9 show that the same speaker (i.e. F1 or M1) can be involved in both very beneficial and very detrimental interactions, depending on the listener. The dotted lines show that same pattern for listeners (e.g. JK or AL) with respect to speakers[2].

[2] As stated above (2.2), subjects participating to the perception task reported normal hearing. A thorough exploration of listener-specific patterns would require ruling out hearing problems through audiometric tests. This was not possible for this study, since the materials used here were collected for independent studies, and subjects were not available for further testing. We could however perform a full audiometric test (using an Amplaid 200 audiometer) on a single listener involved in one of the crucial interactions discussed above (listener JK), and observed normal hearing for all frequency bands (i.e., no hearing loss above the 15dB threshold).

4. Discussion

4.1. Summary of findings

Our results on the encoding and decoding of focus structures in German show the existence of *interacting speaker- and listener-specific strategies*. Specifically,

(i) There is variation among speakers with respect to how phonetic cues are used to encode focus structures, both in terms of robustness (i.e. how many cues are employed) and partitioning (i.e. how many contrasts are expressed by a single cue); cf. 3.1, Table 1.

(ii) Such variation makes the productions of some speakers more intelligible overall than the productions of other speakers; cf. 3.2, Figure 6.

(iii) Listeners vary with respect to how reliable they are in correctly identifying focus structures as intended by speakers; cf. 3.2, Figure 7.

(iv) On top of the overall trends in (ii) and (iii), we document an interaction between individual-specific strategies in production and perception. The same speaker can be more intelligible than average for one particular listener and less intelligible than average for another particular listener; cf. 3.3, Figure 8.

Since subjects do not serve as both speakers and listeners in our dataset, we could not directly verify the hypothesis of a link between accuracy in production and precision in perception, as tested for example by Perkell et al. (2004a,b). However, the result (iv) above seems to question the possibility of understanding accuracy (of production and perception) in absolute terms, that is observing speakers or listeners individually, outside their dyadic interactions.

4.2. Implications for linguistic theory

Apart from being relevant to intonation research, in which the interaction between speaker- and listener-specific behaviour is scarcely documented, we believe the findings above to be of interest to linguistic theory in general.

First, our findings provide additional evidence supporting claims of a complex relationship between phonetic exponents and phonological contrasts. Multiple cues are involved in the signalling of phonological categories, not only in the segmental domain (e.g. Lisker 1986, Coleman 2003) but also for intonational contrasts (see also Cangemi & Grice, submitted). Certain cues, such as voice onset time for voicing contrasts, or peak alignment for pitch accent type contrasts, might be particularly important in both production and perception. However, since they are weighted with respect to other (potentially underexplored) cues, they cannot be treated as the sole exponents of phonological contrasts.

Crucially, the weights associated with phonetic cues in the encoding (and decoding) of contrasts can differ across speakers (and listeners). Even domains in which individual specificity in cue weighting is largely underexplored, as is the case in intonation research, are starting to acknowledge the possibility that (groups of) speakers might encode a phonological contrast by relying more or less strongly on different cues. The study by Niebuhr et al. (2011) mentioned above (1.1) provided initial evidence in this sense, by showing that speakers of Standard Northern German express the contrast between H+L* and H* by primarily varying either peak alignment or peak shape.

Third, acknowledging the interaction between speaker- and listener-specific behaviours leads to a refined understanding of intelligibility (and proficiency) in the encoding (and decoding) of contrasts. Our findings provide evidence that, on top of overall average individual skills as encoders, the intelligibility of individual speakers is modulated by the specificities of the individuals acting as decoders. Likewise, the performance of individual listeners is affected by the specificities of the individual speakers. Our results thus call into question the metaphor "universal donors and recipients" in speech.

Moreover, the results presented here point to the necessity of exploring the cascading of individual specificity in production and perception - something along the lines of the empedoclean gnoseological principle of “like is known by like”. The specific hypothesis to test would be whether an individual X, who *as a listener* has an advantage in decoding a given contrast as produced by speaker Y (rather than by speaker Z), also happens to encode the same contrast *as a speaker* by weighting cues as Y does (rather than as Z does).[3]

Finally, and perhaps most importantly, our findings strengthen the case for the impossibility of conceptualizing phonological categories in the monothetic sense. Rather than singly necessary and jointly sufficient features for category membership, phonetic cues are better understood as dimensions along which phonological categories cluster, in an individual-specific network of phonological knowledge.

Acknowledgements

The work of the first author is funded by the German Research Foundation Excellence Initiative at the University of Cologne (Emerging Group: Dynamic Structuring in Language and Communication). We would like to thank two anonymous reviewers for their useful comments and Roger Mundry (Max Planck Institute for Evolutionary Anthropology, Leipzig) for support with the statistical analysis of speaker-listener interaction. Earlier versions of this study received useful feedback at the 4th Summer School on Speech Production and Perception - Speaker-specific behavior (Aix-en-Provence, October 2013) and at the 14th Conference on Laboratory Phonology (Tachikawa, July 2014).

References

Allen J.S., J.L. Miller, D. DeSteno (2003) Individual talker differences in voice-onset-time. *Journal of the Acoustical Soc403iecty of America* 113(1), 544-552.

Andreeva, B., W. Barry (2007). Cross-language and individual differences in the production and perception of syllabic prominence. *3rd annual meeting of the DFG-Priority Programme 1234*, Cologne, Germany.

Baayen, R.H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Boersma, P., D. Weenink (2010). *Praat: doing phonetics by computer*. Computer program.

Bradlow, A., G. Torretta, D. Pisoni (1996). Intelligibility of normal speech: I. Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication* 20, 255-272.

Byrd, D. (2000). Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica* 57, 3–16.

Cangemi, F., M. Grice (submitted). A distributional approach to categoriality in intonation transcription.

Cassidy, S., J. Harrington (2001). Multi-level annotation in the EMU speech database management system. *Speech Communication* 33, 611–677.

Coleman, J. (2003). Discovering the acoustical correlates of phonological contrasts. *Journal of Phonetics* 31, 351-372.

Grice, M., S. Baumann, R. Benzmüller (2005). German intonation in autosegmental–metrical phonology. In: Jun, S. (ed.), *Prosodic typology: The phonology of intonation and phrasing*, 55–83. Oxford: Oxford University Press.

Grice, M., S. Ritter, H. Niemann, & T. Roettger (ms). *Continuous measures provide insights into the nature of intonation as a marker of focus type*, Unpublished manuscript. University of Cologne, Germany.

Hazan, V., R. Baker (2011). Is consonant perception linked to within-category dispersion or across-category distance? In: *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, 839-842.

[3] Research on this topic is still in its infancy; Idemaru et al. (2012, Exp. 4) failed to provide positive evidence for a within-speaker correlation between production and perception cue weights for stop length contrasts in Japanese.

Hazan, V., R. Romeo, M. Pettinato (2013). The impact of variation in phoneme category structure on consonant intelligibility. *Proceedings ICA Montreal* 19, 1-6.

Idemaru, K., L. Holt, H. Seltman (2012). Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *Journal of the Acoustical Sociecty of America* 132(6), 3950-3964.

Lakoff, G. (1987). Women, fire and dangerous things: What categories reveal about the mind. Chicago and London: The University of Chicago Press.

Lisker, L., A. Abramson (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20(3), 527-565.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and speech* 29(1), 3-11.

Mücke, D., M. Grice (2014). The effect of focus marking on supra-laryngeal articulation – is it mediated by accentuation? *Journal of Phonetics* 44, 47-61.

Newman, R., S. Clouse, J. Burnham (2001). Perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America* 109, 1181-1196.

Niebuhr, O., M. D'Imperio, B. Gili Fivela, F. Cangemi (2011). Are there "shapers" and "aligners"? Individual differences in signaling pitch accent category. *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, 120-123.

Perception Research Systems (2007). *Paradigm Stimulus Presentation*. Computer program.

Perkell, J., F. Guenther, H. Lane, M. Matthies, E. Stockmann, M. Tiede, M. Zandipour (2004a). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *Journal of the Acoustical Society of America* 116, 2338-2344.

Perkell, J., M. Matthies, M. Tiede, H. Lane, M. Zandipour, N. Marrone, E. Stockmann, F. Guenther (2004b). The distinctness of speakers' /s/-/ʃ/ contrast is related to their auditory discrimination and use of an articulatory saturation effect. *Journal of Speech, Language and Hearing Research* 47, 1259-1269.

Schultz, A., A. Francis, F. Llanos (2012). Differential cue weighting in perception and production of consonant voicing. *Journal of the Acoustical Society of America* 132(2), EL95-101.

Jun, S. (ed.) (2014). *Prosodic typology II: The phonology of intonation and phrasing*. Oxford: Oxford University Press.