INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

The Social Life of Setswana Ejectives

Daniel Duran[1], *Jagoda Bruni*[1], *Grzegorz Dogil*[1], *Justus Roux*[2]

[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]South African Centre for Digital Language Resources, South Africa
⟨firstname.lastname⟩@ims.uni-stuttgart.de; justus.roux@nwu.ac.za

Abstract

This paper presents a first phonetic analysis of voiced, devoiced and ejectivized stop sounds in Setswana taken from two different speech databases. It is observed that rules governing the voicing/devoicing processes depend on sociophonetic and ethnolinguistic factors. Speakers, especially women, from the rural North West area of South Africa tend to preserve the phonologically stronger devoiced (or even ejectivized) forms, both in single standing plosives as well as in the post-nasal context (N̥C̥). On the other hand, in the more industrialized area of Gauteng, voiced forms of plosives prevail. The empirically observed data is modelled with *KaMoso*, a computational multi-agent simulation framework. So far, this framework focused on open social structures (*whole world networks*) that facilitate language modernization through exchange between different phonetic forms. The updated model has been enriched with social/phonetic simulation scenarios in which speech agents interact between each other in a so-called *parochial* setting, reflecting smaller, closed communities. Both configurations correspond to the sociopolitical changes that have been taking place in South Africa over the last decades, showing the differences in speech between women and men from rural and industrialized areas of the country.

Index Terms: Setswana, ejectives, stop devoicing, simulation

1. Introduction

South Africa is a very dynamic country, from the ethnological, social and also the linguistic point of view. Over the last decades, changes in the political image of this region influenced naturally all spoken languages and dialects in this area. The current eleven official languages are now coexisting and interacting in every-day speech, providing cross-language exchanges and influencing forms on different linguistic levels [1]. Thus, a broader, sociolinguistic perspective in studying linguistic changes of this region seems appropriate. The importance of this type of analysis in the variation of speech has also been illustrated by Foulkes and Docherty [2], who point out that different social, age and gender groups tend to use diverse phonetic and phonological forms of segments' variation (e.g. using different laryngealization forms).

A change of this kind can be observed in Setswana plosives, on the phonetic/phonological level. As other languages from the Sotho-Tswana group (e.g. Sesotho sa Leboa and Sesotho), Setswana contains nasal-stop clusters (NC) in which a devoicing of the post-nasal plosive (N̥C̥) is a common process [3, 4, 5, 6]. It is said to have appeared in Bantu languages in order to facilitate production of voicing during the stop segment, which was lost later during language evolutionary changes. This process has been described as articulatorily counter intuitive [7], in that voicelessness requires additional articulatory cost, whereas voicing reflects a neutral state in a post-nasal position. Similarly from the phonological perspective, Pater [3] accounts for a *N̥C̥ constraint, claiming that many languages demonstrate existence of prenasalized voiced stops but lack prenasalized voiceless stops. The rule penalizes consonantal sequences of [+nasal] followed by [-voice] and Pater [3] claims that N̥C̥ clusters seem to be uncommon in a variety of languages. He states that typological data, as well as phonetic evidence argue for a universal but violable *N̥C̥ constraint.

This paper presents a phonetic analysis of voiced, devoiced and ejectivized stop sounds in Setswana. The present approach investigates apparent phonetic and phonological sound changes in Setswana by combining the acoustic analysis from two speech databases [5, 8] and a new model of the social setting (the so-called *parochial* network, cf. [9]). Our acoustic analysis demonstrates that the Setswana plosives /p, t, k/ in post-nasal contexts as well as singletons can obtain several voicing statuses. Starting from fully voiced, through devoiced, up to 'hyper' devoiced ejective-forms. The acoustic analysis of both corpora [5, 8] shows that devoiced variants prevail in Setswana. Our analysis is based on previous classification work by Kingston [10] and Fallon [11]. Following this earlier work, we assess acoustic parameters to define the difference between devoiced and ejectivized stops (see Table 1). The analysis of the NCHLT corpus [8] shows that ejectivization, i.e. phonological strengthening of devoiced stops, occurs more often in the rural areas of the country, mostly among women who spend their time in closed tribal communities, described in this paper as *parishes* (see below). Our current computational simulation serves as an explanation to which extent the different social networks play a role in the formation of voicing forms in the current language, demonstrating differences in variant competition within different social network settings.

In the following section we first briefly describe the speech data and the acoustic analysis. Section 3 describes the computational model and after that we present our results.

2. Acoustic analysis

2.1. The Coetzee and Pretorius data

The first corpus data that we examined was collected and analysed by Coetzee and Pretorius [5]. It contains speech of 12 native speakers of Setswana (born in the area around Potchefstroom) from an academic environment. The data is fully annotated and consists of carrier phrases with real and nonsense words constructed to analyse plosives in a post-nasal context.

2.2. The NCHLT corpus

The NCHLT Setswana Speech Corpus (National Centre for Human Language Technology; [8]) was originally created for the development of Automatic Speech Recognition systems for South African languages. It contains speech material for all

http://dx.doi.org/10.21437/Interspeech.2017-922

Table 1: *Acoustic classification of ejectives (updated from Kingston [10])*

Acoustic cue	Fortis ejective	Lenis ejective
F0 of a following vowel	elevated	depressed
VOT	long	short
Type of burst	intense, silence	weak
Transition to the max. amplitude	abrupt, rapid rise	gradual, slow
Type of voice at VOT	modal	creaky
Vocal folds	stiff	slack

eleven languages in South Africa, among it Setswana with 210 speakers (in total 109 female, 101 male, 56:19h of speech). The data consists of carefully enunciated read speech (e.g. from South African government websites) and was collected via a smartphone Application (*Woefzela*). It is available online at http://rma.nwu.ac.za. We report here on data from 94 male and female speakers in four age groups (below 20, 20–30, 31–40 and 41 plus) from two geographical regions (North West and Gauteng).

2.3. Acoustic analysis of ejectives in the nasal and non-nasal contexts (NCHLT corpus)

In the analysis of the NCHLT data we have mainly focused on the processes of phonological strengthening, i.e. on the observation of the amount of devoiced stops and ejectives. According to the system of acoustic cues proposed by Kingston [10] (as shown in Table 1), the ejectives from the NCHLT corpus were selected using a custom implementation of an automatized Praat [12] plug-in[1] function to identify candidate segments within the corpus which were then labelled and classified manually.

We distinguish three different types of stop segments: fully voiced, devoiced, and ejective or ejectivized ('hyper' devoiced) forms. The key factor in the distinction was the f0 contour in the vowel following the plosive, as well as VOT length (for long: >80 ms, for short: <20 ms), type of burst and the type of transition to the maximal amplitude. Finally, the type of modality (modal vs. creak) was also a determiner in the decision on the presence of ejectivization. We also place these phonetic characteristics on a phonological scale where presence of [voice] is considered a weakening of a segment (voicing), whereas [-voice] strengthening of a segment (ejectivization).

3. Computational simulation

We model the observed linguistic situation of Setswana computationally using the *KaMoso* simulation framework [13][2]. Within this dynamic multi-agent simulation model, a population of speaker-listeners interacts in different social networks. The simulation goes through a number of *epochs* modelling the dynamic evolution of the population over time. In each epoch, the agents receive input from other agents ("teachers"). The model employs exemplar-theoretic principles, assuming that linguistic categories are mentally represented by collections of previously encountered exemplars [14, 15, 16, 17]. In this framework, each

[1]Kindly provided by Jörg Mayer (personal communication).

[2]The Java source code as well as the configurations used for the simulations presented here are available online at https://github.com/simphon/KaMoso

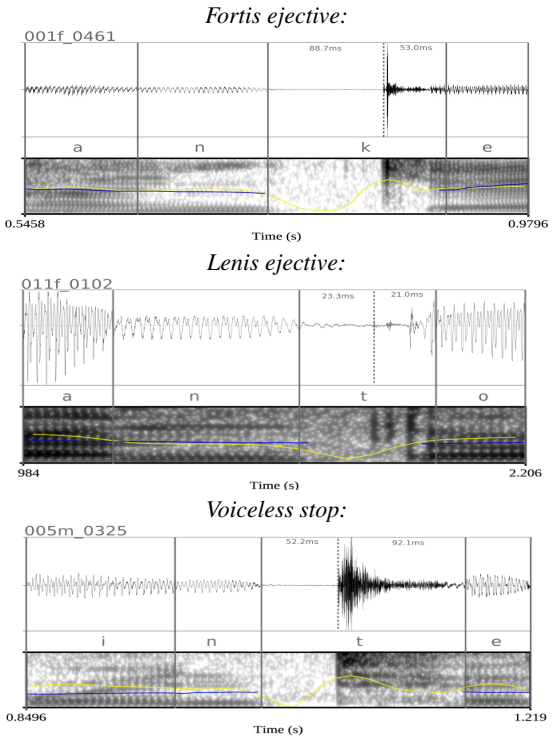

Figure 1: *Examples of fortis and lenis ejectives and voiceless stop from the NCHLT corpus, labelled [k], [t] and [t], respectively. The yellow curve on the spectrogram indicates intensity, and the dark blue line pitch.*

perceived speech item is stored in memory retaining its full phonetic detail as well as indexical information like speaker identity. We refer to the set of all stored exemplars as the *lexicon*. Production of speech items is based on a speaker's lexicon with each exemplar serving as a potential production target according to its weight (or "activation"). Noise is added to exemplars in production. Before perceived exemplars are stored in memory, they are warped slightly towards local distribution maxima (implementing a *perceptual magnet effect*, cf. [18, 19]).

The model describes the competition between two phonetic variants A and B over a large number of epochs. Given free variation (as in the case of the Setswana stops), weighting of exemplars may take into account different features: the *phonetic prototypicality* of an exemplar (i.e. its similarity to the category centroid), or non-linguistic social features like the *social closeness* of the exemplar's original speaker or the *social status* of the speaker [20]. In our simulations, two interaction schemes are compared: *closeness interaction* (favouring variants from socially close individuals), and *status interaction* (favouring variants from socially influential individuals).

In order to model an influence of the speakers' regional backgrounds on their phonetic production patterns (as observed in the corpus data, see Section 4.1 below), we investigate different network topologies. Formally, the social networks are represented by undirected graphs, where nodes correspond to individuals (or agents) and edges between nodes correspond to direct social relations. Social distance can then be defined as the minimum number of edges between any two nodes (the minimal path length); and social *closeness* as an inverse of this distance measure. As a representation of the social structure with largely

closed language communities, we use a network with a number of clusters or *parishes*. Each parish is connected to other parishes by a (randomly chosen) link, such that all parishes together form a connected network graph (i.e. there exists a path between any given pair of agents). Within each parish, the network graph has a local *small world* topology [21, 22] which is supposed to be a good formal representation of real social network structures. This parochial configuration has the effect that the average path length within each parish is relatively small, but across parish boundaries the average path lengths are relatively large. We refer to this setting as the *parochial network*. This is contrasted with a simple *small-world network* (an alternative name could also be *whole world network* [9]) where social distance between any two agents is relatively small. We assume that this latter type of network corresponds to a socially open (urban) community.

We run each simulation over a number of epochs. We also run the simulations repeatedly with the same sets of parameters as each individual simulation run is subject to random variations introduced at various points in the model. For the simulations reported here, we define a population of N = 400 agents. One fixed set of agents has been generated with randomly assigned social status in the range of 0.01–0.25, and a small number of hyper-influential agents with status 1.0 (cf. [20]), randomly assigned gender (50% female), and equally distributed ages 1–5. The initial ratio of A-variant tokens in the agents' lexicons is 0.95 and 0.85 for female and male agents, respectively. For hyper-influential agents, the corresponding values are 0.05 and 0.15. The lifespan of an agent is 5. When this age is reached, that respective agent is replaced by a new-born agent which is generated according to the same principles (with an initially empty lexicon).

The initial set of agents is used for all simulations in order to focus on the effects of the different network topologies which just change the connections between agents.

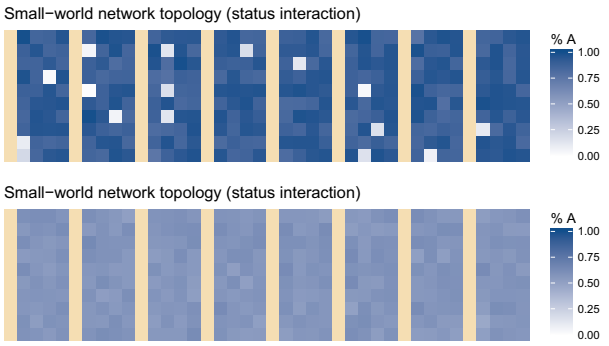

Figure 2: *Distribution of A-variant exemplars in the lexicons of all agents in the population. Each coloured square tile represents one agent. Cream-coloured tiles corresponds to empty lexicons of new-born agents. Top panel: state at beginning of simulation. Bottom panel: state at the beginning of epoch 800.*

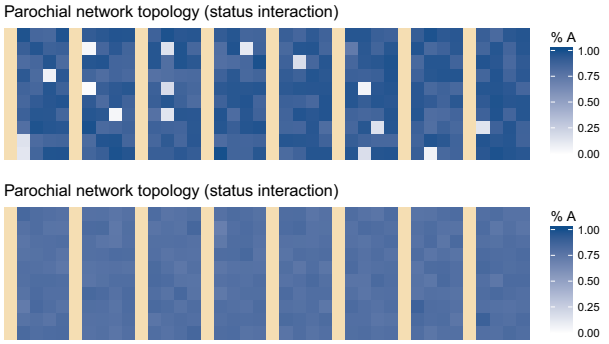

Figure 3: *Distribution of A-variant exemplars in the lexicons of all agents in the population. Each coloured tile represents one agent. Cream-coloured tiles corresponds to empty lexicons of new-born agents. Top panel: state at beginning of simulation. Bottom panel: state at the beginning of epoch 800.*

4. Results

4.1. Acoustic analysis of ejectives

Based on the acoustic analysis of their database, Coetzee and Pretorius [5] divided the speakers into two varieties: (1.) speakers who produce more than 90% of /m+bV/-sequences as [mpV] and (2.) speakers who show a tendency for post-nasal voicing. Overall it has been demonstrated that 80% out of all analysed stops devoice post-nasally and that less than 20% of post-nasal stops were articulated as ejectives.

For the post-nasal ejectives, we analysed samples from 309 sentences from the NCHTL corpus (93 speakers), and for singleton ejectives, we analysed samples from 512 sentences (16 speakers). In this data, in the post-nasal context 39% of stops were classified as voiceless and 6% as ejectives (see Figure 1 for prototypical examples from the corpus). Out of this 6%, the majority (72%) occurred in the velar context VnkV, (11% in the VntV and 17% in the VnpV context). Female speakers articulated more ejectives (9%) than males (3%), where as much as 80% of females produced at least one ejective. More of the ejective stops occurred in the speech from the rural North West area (7% out of all stops analysed) and 3% in the industrialized Gauteng area. In the North West area 80% of speakers produced at least one ejective. The age distribution in production of ejectives shows that most of them were produced by speakers between the 20th and 30th year of life. Out of all analysed stops, voiceless stops in the post-nasal context were produced by 64% of female speakers and 53% of males. In this classification 53% were produced in the speech from the North West area and 34% from the Gauteng area. Thus, it is reasonable to state that in the post-nasal context, voiceless stops behaved similarly to ejectives in terms of their geographical distribution and across genders.

In the non-nasal context out of all stops analysed, 51% were classified as ejectives, out of which 20% were so called 'fortis ejectives' (see Table 1 with the acoustic key parameters). Most of them (73%) were produced by male speakers and the majority (61%) in the Gauteng area occurred in the velar context (VkV). In the classification of the voiceless stops it has been observed that 52% were produced in the Gauteng area and the majority was found in the speech of the female speakers (57%). Most of the voiceless stops found in the data were bilabial plosives (VpV). In this view, the group of voiceless stops in the non-nasal context behaves rather differently than the ejectives in the non-nasal context. Rather, the classification shows that regarding gender and geographical area, voiceless stops in post-nasal and non-nasal context behave similarly.

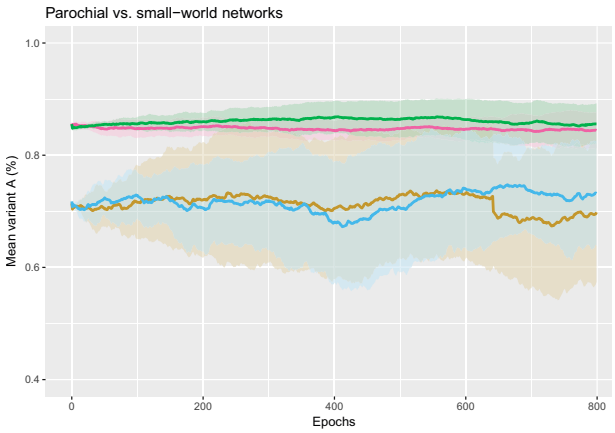

Figure 4: *Mean A-variant ratio in 11 repetitions of simulations: parochial with closeness interaction; parochial with status interaction; small-world with closeness interaction; small-world with status interaction. Solid lines indicate mean values, shaded areas mean ± standard deviation.*

4.2. Computational simulation model

In order to visualize the distributions of variants A and the initially rare variant B across the lexicons of all agents from the population, we count the exemplars in the agents' lexicons and show the ratio of A-variant exemplars in a heatmap-like figure where each square represents one agent: Figure 2 shows the distribution of variant A from a simulation on a small-world network with an interaction scheme based on social closeness. Darker shades of blue indicate a higher ratio of A-variant exemplars in an agent's lexicon. New-born agents with empty lexicons are indicated by pale yellow squares, which are visible as stripes in the plot due to the regular age distributions of the population and the graphical arrangement of the nodes. Note that the horizontal and vertical axes have no meaning. The same is true for the actual position of a node in space. The figure shows the initial state (epoch 0) in the top panel and below the state at the beginning of epoch 800. Figure 3 shows the distribution of variant A for a simulation with the same parameters on a parochial network. Note that the initial state is the same in Figures 2 and 3 as the same set of agents has been used to initialize the network.

The results indicate that the small-world network topology seems to allow for the initially rare variant to spread through the community (indicated by the overall lower percentages of the A variant).

In order to illustrate the evolution of the variant competition over time, we count the number of produced tokens in each epoch and determine which belong to either variant. Figure 4 shows the mean ratio of produced A-variant exemplars for the four different simulation setups (*parochial* vs. *small-world* network with *closeness* vs. *status-based* interactions). The mean is computed of 11 repeated runs with the same initial parameters for each of the four different setups. As this figure shows, the distribution remains relatively stable after the first exchanges (note that the initial state of the population is always the same, but the starting points of the curves differ due to the fact that the values are computed at the end of an epoch after all interactions have been carried out).

5. Discussion and Conclusions

Our simulations indicate that a parochial social network structure may favour preservation of phonetically and phonologically 'conservative' forms. Sociolinguistic studies describing the South African language politics like [1] indicate complexity of multilingualism and heteroglossia in this area. Ours is an attempt to demonstrate change within one of the official languages, i.e. Setswana, where the general social pattern of the country is reflected. The phonologically stronger but phonetically and articulatory less intuitive devoiced and ejectivized variants of post-nasal stops dominate in the simulation within the closed community (*parochial*) scenario. Reflected partially in our acoustic measurements, it is also in line with the hypothesis that the protolanguage forms of the Bantu group (*Ur-Bantu*, [23]) are better preserved when 'isolated' from the influences of other official South African languages and/or dialects. The use of computational phonetic simulation [20, 24, 25, 19] as a means of investigation in the study of language change is confirmed in that the integration of real-life existing sociophonetic factors brings about the current phonetic shape of Setswana voicing patterning.

6. Acknowledgements

This work is funded by the German Research Foundation (DFG) within the Collaborative Research Center SFB 732 / A2. We would like to thank Andries W. Coetzee for kindly sharing his data and for discussions on issues of Setswana stop (de)voicing. Many thanks to Hanna Kicherer for help in the acoustic analysis.

7. References

[1] L. Hibbert, *The linguistic landscape of Post-Apartheid South Africa: politics and discourse*. Bristol ; Buffalo: Multilingual Matters, 2016.

[2] P. Foulkes and G. Docherty, "The social life of phonetics and phonology," *Journal of Phonetics*, vol. 34, no. 4, pp. 409–438, 2006.

[3] J. Pater, "Austronesian nasal substitution and other NC̥ effects," in *The Prosody-Morphology interface*, R. Kager, H. van der Hulst, and W. Zonnenveld, Eds. Cambridge: Cambridge University Press, 1999, pp. 310–343.

[4] A. W. Coetzee, S. Lin, and R. Pretorius, "Post-nasal devoicing in Tswana," in *Proceesdings of ICPhS XVI*, Saarbrücken, Aug. 2007, pp. 861–864.

[5] A. W. Coetzee and R. Pretorius, "Phonetically grounded phonology and sound change: The case of Tswana labial plosives," *J. of Phonetics*, vol. 38, no. 3, pp. 404–421, 2010.

[6] M. Gouskova, E. Zsiga, and O. T. Boyer, "Grounded constraints and the consonants of Setswana," *Lingua*, vol. 121, no. 15, pp. 2120–2152, 2011.

[7] P. A. Keating, "Underspecification in phonetics," *Phonology*, vol. 5, pp. 275–292, 8 1988.

[8] E. Barnard, M. H. Davel, C. van Heerden, F. de Wet, and J. Badenhorst, "The NCHLT Speech Corpus of the South African languages," in *Proc. of SLTU*, St. Petersburg, Russia, 2014, pp. 194–200.

[9] G. Dogil, J. Bruni, D. Duran, J. Roux, and A. W. Coetzee, "Social dynamics and phonological strength: Post-nasal devoicing in Tswana," in *LabPhon15: Speech Dynamics and Phonological Representation*, Cornell University, USA, 2016, p. Abstract 125.

[10] J. C. Kingston, "The phonetics and phonology of the timing of oral and glottal events," Doctoral Dissertation, University of California, Berkeley, 1985.

[11] P. D. Fallon, *The Synchronic and Diachronic Phonology of Ejectives*. Hoboken: Taylor and Francis, 2013.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Mar. 2017, version 6.0.24. [Online]. Available: http://www.praat.org/

[13] D. Duran, J. Bruni, M. Walsh, and G. Dogil, "A hybrid model to investigate language change," in *Proceedings of ICPhS*, 2015, paper 735.

[14] S. D. Goldinger, "Words and voices – perception and production in an episodic lexicon," in *Talker Variability in Speech Processing*, K. Johnson and J. Mullennix, Eds. Academic Press, 1997, ch. 3, pp. 33–66.

[15] K. Johnson, "Speech perception without speaker normalization: An exemplar model," in *Talker Variability in Speech Processing*, K. Johnson and J. Mullennix, Eds. Academic Press, 1997, pp. 145–165.

[16] J. B. Pierrehumbert, "Exemplar dynamics: Word frequency, lenition, and contrast," in *Frequency and the Emergence of Linguistic Structure*, J. L. Bybee and P. Hopper, Eds. John Benjamins Publishing, 2001, pp. 137–157.

[17] A. Wedel, "Category competition drives contrast maintenance within an exemplar-based production/perception loop," in *Proc. 7th Meeting of the ACL SIG Computational Phonology*, 2004, pp. 1–10.

[18] P. K. Kuhl, "Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not," *Perception & Psychophysics*, vol. 50, no. 2, pp. 93–107, 1991.

[19] A. Wedel, "Exemplar models, evolution and language change," *The Linguistic Review*, vol. 23, pp. 247–274, 2006.

[20] D. Nettle, "Using social impact theory to simulate language change," *Lingua*, vol. 108, no. 2-3, pp. 95–117, 1999.

[21] S. Milgram, "The small-world problem," *Psychology Today*, vol. 1, no. 1, pp. 61–67, 1967.

[22] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[23] C. Meinhof, *Introduction to the phonology of the Bantu languages: being the English version of "Grundriss einer Lautlehre der Bantusprachen"*. Berlin; London: Dietrich Reimer (Ernst Vohsen); Williams & Norgate, Ltd., 1932, ed. and trans. by Nicolaas Jacobus van Warmelo.

[24] A. Baker, "Computational approaches to the study of language change," *Language and Linguistics Compass*, vol. 2, no. 2, pp. 289–307, 2008.

[25] P. Boersma and S. Hamann, "The evolution of auditory dispersion in bidirectional constraint grammars," *Phonology*, vol. 25, pp. 217–270, 2008.

[26] K. Johnson and J. Mullennix, Eds., *Talker Variability in Speech Processing*. Academic Press, 1997.