Acoustics'08 Paris
June 29-July 4, 2008
www.acoustics08-paris.org

An acoustic investigation into coarticulation and speech motor control: high vs. low frequency syllables

Frank Herrmann, Sandra Whiteside and Stuart Cunningham

University of Sheffield, Human Communication Sciences, 31 Claremont Crescent, S10 2TA
Sheffield, UK
f.herrmann@sheffield.ac.uk

Psycholinguistic research of the mid nineties suggests that articulatory routines for high frequency syllables are stored in the form of gestural scores in a library [5, 8]. Syllable frequency effects on naming latency and utterance duration have been interpreted as supporting evidence for such a *syllabary* [3].

This paper presents a data-subset from a project investigating speech motor learning as a function of syllable type. Fourteen native speakers of English were asked to listen to and repeat 16 mono-syllabic stimuli which belonged to either of two categories: high & low frequency syllables (CELEX).

Acoustic coarticulation measures, i.e. F2 locus equations and absolute formant changes, were used to indirectly determine the degree of gestural overlap in articulatory movements. In addition, utterance durations were measured to determine speed of articulation. Significant syllable frequency effects were found for both **F2 Locus equations** (e.g. slope and R²), and **utterance duration**. High frequency syllables exhibited greater degrees of coarticulation (steeper slopes), greater overall consistency in their production (greater R²) and shorter utterance durations than low frequency syllables. These data provide some further supporting evidence that different syllable categories may be encoded differently during speech production.

1 Introduction

During speech production, around 60 muscle groups cooperate to produce up to ten or even as much as fifteen sounds per second; thus, posing highly complex computational demands with large degrees of freedom to our speech production system [1]. Schiller *et al.* (1996) showed that speakers re-use a rather small set of syllables even though languages such as English or German may have more than 12,000 different syllables [2]. Cholin *et al.* (2006) suggest it would be more efficient to store the verbo-motor patterns of these frequently used syllables and retrieve them as wholes instead of computing them anew each time they appear during syllabification [3]. This way, the degrees of freedom and the computational load could be considerably reduced during speech encoding.

This concept goes back to Crompton (1982) [4] and later Levelt and Wheeldon (1994) who propose the notion of the **mental syllabary** [5], a repository of articulatory-phonetic syllable programs in form of gestural scores [6], which serve as basis for the motor execution of high frequency syllables. Significant syllable frequency effects on the naming latency and utterance duration of nonsense words were interpreted as supporting evidence for the existence of such a syllabary [3, 5].

In analogy to dual-route reading, Whiteside and Varley (1998a, b) proposed that there may be two possible phonetic encoding routes employed during speech production: a **direct** and an **indirect route** [8, 9, 7].

Similar to Levelt and Wheeldon's concept, the direct route makes greater use of stored schemas and operates for high frequency syllables; yet with frequency being the key element to the direct route they further argued that any unit of speech, even above the syllable level, can be potentially stored as a schema. The indirect route relies more heavily on online computation of sub-syllabic units and is employed for low frequency and novel syllables (see Figure 1). This implies **more rapid** and **less error prone** productions for the sequences encoded via the direct route. Whiteside and Varley took this one step further and suggested that direct route encoding would result in **greater consistency** and **cohesiveness** of the speech productions [9]. The latter of which can be indirectly determined acoustically by analysing the degree of gestural overlap, i.e. coarticulation.

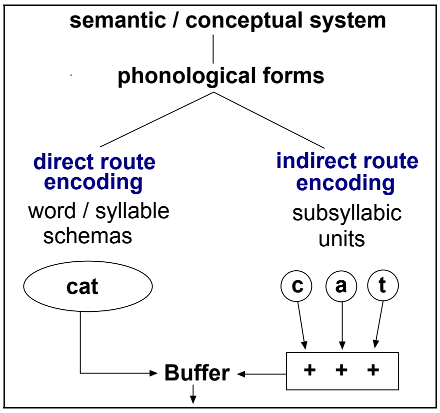

Figure 1: Dual-route phonetic encoding (adapted from Whiteside and Varley, 2001 [7]).

In a pilot study (Whiteside and Varley, 1998a), seven healthy young speakers (mean age 25.4 years) of Standard Southern British English repeated randomised high and low frequency monosyllabic stimuli after an experimenter [8]. The stimuli were preceded by either 'a' or 'the'. Durational measures were response latency, utterance, word, and formant transition duration. Coarticulation was quantified by calculating the F2 changes between vowel onset and temporal midpoint.

Based on an analysis of the recordings of the first repetition, the suggestion that naming latency is a function of syllable frequency [3, 5] was supported. Neither utterance nor word duration showed significant differences between the two syllable categories, but the difference in transition duration was significant. Even though their data set did not show significant differences in F2 changes, it showed a clear trend for F2 changes being smaller, thus indicating greater articulatory overlap, and more consistent (smaller SD values) for high frequency than for low frequency syllables. However, the scope of their study was limited by the small number of participants and analysed repetitions.

The current study, which is part of a larger project on speech motor learning, revisits the concept of the dual-route phonetic encoding model and tries to empirically underpin the trends found in the durational [3, 5] and coarticulation measures [9].

Working hypotheses for the presented data are: If there are two different phonetic encoding routes, then high frequency syllables should exhibit **shorter utterance durations**, **greater degrees of coarticulation**, and **greater overall consistency** than low frequency syllables [3, 5, 7, 8].

Previous literature suggests that the degree of coarticulation is influenced by speaker sex. This was found when quantifying coarticulation using absolute formant changes [10], yet did not show in F2 locus equations [11]. We would expect syllable frequency effects to be independent of speaker sex effects.

2 Method

2.1 Data collection

This data set is part of a larger project on speech motor learning, which comprises of 24 mono- and 24 disyllables belonging to three categories: high & low frequency and exotic (i.e. containing non-English sounds). The subset presented here, was collected at the entry point of the study and contains the **low** and **high frequency monosyllables** as found in the CELEX database [12]; mean values for the two categories are 324.5 (high) and 3.25 (low) occurrences per million, respectively.

The stimuli were matched for phonetic complexity; their intrinsic stimulus duration as well as vowel quality were also considered in as far as possible. Examples for each category are /biz/ (356 / million) versus /bis/ (6 / million).

Seven male and seven female native speakers of English took part in this study (mean age: 19;8 years). None of the participants had a known speech or language impairment and all underwent a screening process comprising of a hearing test (20dB HL) as well as an interview to elucidate any medication usage and to collect other demographic details such as accent.

The recordings were conducted in a sound attenuating booth to secure recordings of high quality and to avoid any distractions from the mirroring task the participants were asked to perform. The auditory stimuli were presented in randomised order, yet ten repetitions of each were recorded before moving on to the next stimulus. Participants would listen to a stimulus, repeat it, listen again, repeat again, and so forth until the tenth repetition; a beep then signalled a new stimulus. Thus, the cognitive load on the participants was reduced in as far as possible and the regular rhythm of the stimulus presentation encouraged the participants to use a consistent prosody, i.e. overall speech rate and intonation pattern.

The recorder used was a *Marantz PMD670* and its settings were: mono at a sampling frequency of 22.05 kHz and a 16bit sampling rate. The microphone was a *Sennheiser MD425*.

2.2 Data analysis

The acoustic analysis software Praat (Version 5.0.19) was used to obtain measurements. TextGrid files were generated to mark points of interest, which were visually determined based on the sound pressure waveform and spectrographic display with overlaid glottal pulses and formant plots. Based on these a script measured the relevant formant frequencies and calculated durational measures based on the time indexes.

The formant analysis parameters for Praat's algorithm were set at a window length of 25ms, a time step of 5ms, and a pre-emphasis from 60Hz. Depending on the speaker sex the ratio of *number of formants / maximum formant frequency* was changed to *5 / 5,500Hz* (female) and *5 / 5,000Hz* (male) as suggested in Praat's user guide.

For **Durational Measures** the points of interest were acoustic onset and offset, which were used to determine the *utterance duration*. The onset and endpoint of the vowel were marked to calculate its temporal midpoint and the formant *transition duration.*

Consonant to Vowel **Coarticulation Measures** were comprised of the first three formants, which were measured at vowel onset position (Fn_{ons}), i.e. in the first glottal pulse, and at the temporal midpoint of the vowel (Fn_{mid}). Similar to Whiteside and Varley's study, coarticulation was then quantified by calculating the *absolute formant changes* (see Eq. 1). The smaller ΔFn is, e.g. ΔF2 [8], the greater is the articulatory overlap.

$$\Delta Fn = abs(Fn_{mid} - Fn_{onset}) \quad (1)$$

In addition, *F2 locus equations*, which have been previously identified as being indicative of the degree of coarticulation [13], were derived (see Eq. 2) for the high and low frequency syllable data.

$$F2_{onset} = y + slope \times F2_{mid} \quad (2)$$

F2 locus equations are linear regression lines which are a result of correlating F2 onset values with F2 midpoint values, i.e. using F2 midpoint as the predictor. Thus, while absolute formant changes, such as ΔF2, deprive the resulting value of its acoustic context, F2 locus equations preserve it [14]. The **slope** of the regression line is indicative of the degree of coarticulation or, for the purposes of this study, cohesiveness; the steeper the slope the greater the articulatory overlap. **R²**, or goodness of fit of the regression line, is indicative of the overall consistency of the data, and thus the consistency of speech production patterns over multiple repetitions.

3 Results

The mean values of the durational measures as well as the absolute formant changes in F1, F2 and F3 per stimulus and participant were calculated.

A mixed ANOVA (repeated measures with a between subjects measure) was used to analyse syllable frequency effects and sex differences on these variables. Bonferroni adjustments for multiple comparisons were implemented.

3.1 Durational measures

The means and standard deviation values for utterance duration and transition duration are given in Table 1 by syllable type. The durational measures showed that the mean difference of 40.01ms in **utterance duration** was significant [$F(1, 99) = 7.57$, $p < .01$]; the mean duration of the high frequency syllables was significantly shorter than that of low frequency syllables. The mean difference for

transition duration of 10.71ms, however, was not significant (p >.05).

N=101	High (SD)	Low (SD)	Mean difference
Utterance duration	597.95 (120.14)	637.96 (152.31)	40.01
Transition duration	111.96 (101.25)	101.25 (34.88)	10.71

Table 1: Mean and standard deviation values for utterance and transition durations (in ms) for High and Low Frequency Syllables

3.2 Coarticulation measures

The means and standard deviations for the **absolute formant changes** for F1, F2 and F3 are given in Table 2. Mean differences in the absolute formant changes between High and Low frequency syllables were not significant for **ΔF1** [F(1, 99) = 0.25, p >.05] and **ΔF2** [F(1, 99) = 1.58, p >.05]. However, data for **ΔF3** approached significance [F(1, 99) = 3.49, p =.07].

N=101	High (SD)	Low (SD)	Mean difference
ΔF1	90.07 (72.68)	85.05 (75.24)	5.02
ΔF2	193.11 (115.60)	211.09 (146.05)	17.98
ΔF3	162.85 (80.16)	191.29 (120.34)	28.45

Table 2: Mean values, standard deviations, and mean differences for the absolute formant changes (in Hz) for High and Low Frequency Syllables.

A comparison of the **F2 locus equations**, based on the raw data for all speakers *pooled together*, showed an overall effect of syllable frequency (see Figure 2).

A Z-test revealed that the **slopes** for the high frequency syllables (.627) and low frequency syllables (.494) were significantly different (Z = 7.78, p <.0001). In addition, the differences between the correlation coefficients underlying the **R² values** for the high (r =.925) and low (r =.723) frequency syllables were tested using Fisher's z_r transformation test. The results showed a significant difference between the correlation coefficients representing the high and low frequency syllables (z=15.97, p <.0001)

The slope, R² and y-intercept values are given in Table 3 for each subject by syllable frequency. In addition, the R² values were plotted against the slope values for the F2 locus equations for each individual participant (see Figure 3).

	High Frequency Syllables			**Low Frequency Syllables**		
	y-intercept	slope	R²	y-intercept	slope	R²
F1	779.34	0.57	0.93	1179.68	0.43	0.46
F2	630.01	0.58	0.87	802.99	0.58	0.55
F3	677.7	0.59	0.88	1091.98	0.41	0.58
F4	806.63	0.55	0.85	1473.65	0.18	0.29
F5	509.05	0.67	0.94	1064.27	0.45	0.49
F6	538.46	0.68	0.88	700.47	0.65	0.56
F7	497.06	0.69	0.94	1357.77	0.33	0.26
M1	705.37	0.56	0.78	839.49	0.53	0.55
M2	444.36	0.69	0.92	770.76	0.55	0.78
M3	599.4	0.59	0.84	1200.49	0.31	0.29
M4	901.93	0.46	0.63	1143.28	0.36	0.45
M5	1002.99	0.43	0.51	845.79	0.51	0.63
M6	480.38	0.66	0.84	1457.06	0.11	0.02
M7	486.06	0.68	0.91	1025.77	0.41	0.66

Table 3: Slope, R² and y-intercept values for all female (Fn) and male (Mn) participants by syllable frequency.

The data in Table 3 and Figure 3 provide further evidence for an effect of syllable frequency. The slope values for the high frequency syllables trend for higher values (.43 to .69) compared to the low frequency syllables (.11 to .65).

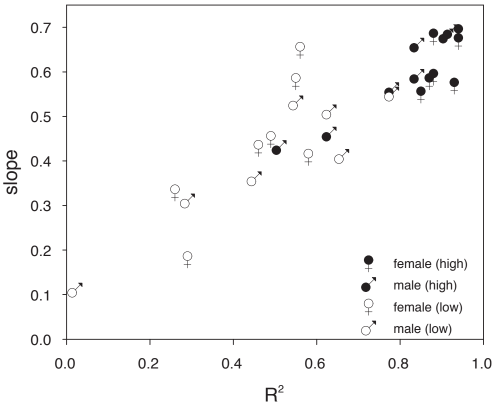

Figure 3: Slope and R² of the F2 locus equations coded by syllable frequency and speaker sex.

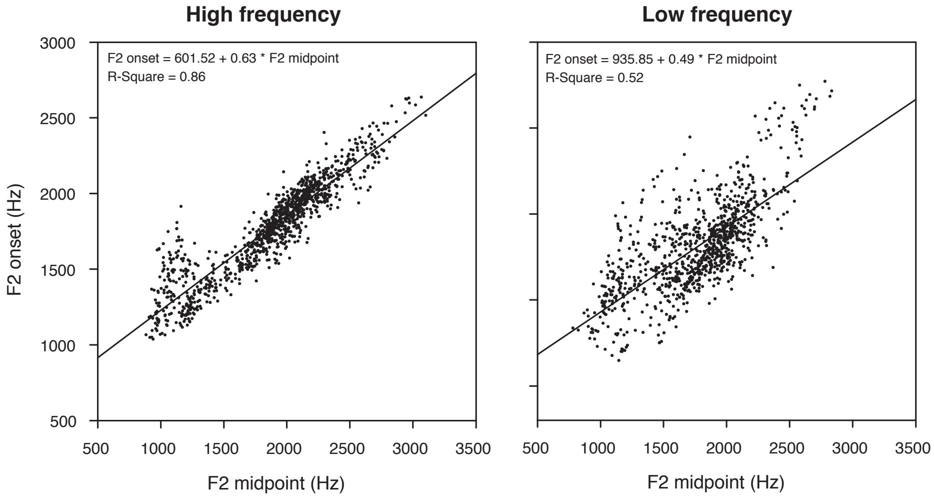

Figure 2: F2 locus equations for all participants across all productions for High and Low Frequency Syllables.

3.3 Sex differences

Neither **utterance duration** [F(1, 99) = 0.16, p =.69], nor **transition duration** [F(1, 99) = 0.43, p =.51] proved to be significantly different.

Figure 4 displays the mean values (+/- 1SD) for the **absolute formant frequency changes** for F1, F2 and F3 by speaker sex. There were significant differences in the absolute formant frequency changes. On average, female speakers had higher ΔF1 values (mean diff. 29.5Hz / 37.5%); [F(1, 99) = 8.03, p <.01], and higher ΔF2 values (mean diff.: 56.9Hz / 31.8%) [F(1, 99) = 7.07, p <.01] than male speakers. The mean difference of 4.7 Hz for ΔF3 was not significant [F(1, 99) = 0.12, p =.73].

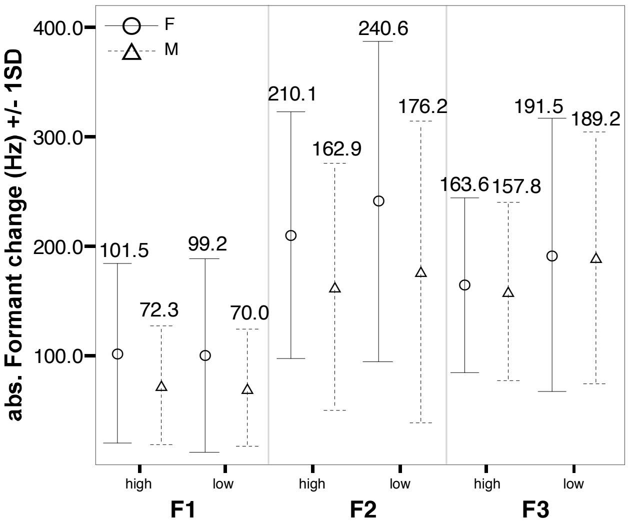

Figure 4: Mean values (+/- 1SD) for ΔF1, ΔF2, and ΔF3 according to speaker sex and syllable frequency.

Figure 4 suggests that the overall pattern is similar for both sexes. Indeed, the syllable ***frequency*** **x** ***sex*** interaction was neither significant for ΔF1 [F(1, 99) = 0.001, p =.97], nor for ΔF2 [F(1, 99) = 0.42, p =.52].

There were no significant effects of speaker sex on **the F2 locus equations** coefficients for either slope [F(1, 12) = 0.59, p >.05] or R² [F(1, 12) = 0.62, p >.05] (see Figure 3). This confirms previous findings by Sussman and colleagues [11].

4 Discussion

Data obtained from 14 speakers suggests that **utterance duration** and **F2 locus equations** vary as a function of syllable frequency; with speaker sex having no significant influence on either of these measures.

The **utterance duration** of high frequency syllables was overall shorter than for low frequency syllables (mean diff. 40.01ms). Also, taking the difference in SD values into account, the overall temporal consistency was greater for high frequency (SD 120.14ms) compared to low frequency syllables (SD 152.31ms). This finding is in contrast to Whiteside and Varley's study, where word duration did not differ significantly [8]. The limited number of recordings which were analysed or the use of a preceding article might have caused this, i.e. the monosyllabic utterance was turned into a phrase, which may have lead to a different planning strategy for the participants.

The opposite was the case for **Transition duration**, which did not prove to be significantly different for the syllable frequency categories. However, formant transition duration per se was not measured in either study – it was rather half the vowel duration. Transition duration would be more clearly quantified if one measured from the onset of the vowel to the minimum/maximum position for each of the individual formants. However, there are not always clear maxima/minima positions identifiable, as the overall formant shapes are largely determined by the specific CVC sequences. It would therefore be interesting to analyse a less arbitrary subset of the data, i.e. those sequences showing clear formant minima/maxima, to see if there are trends in formant transition duration.

Out of the *two* **coarticulation measures**, i.e. absolute formant changes and F2 Locus equations, only the latter has proven to be a function of syllable frequency. This seemingly contradictory finding, i.e. ΔF2 vs. F2 locus equation, needs further discussion. For the two-way mixed ANOVA, the *means* of the absolute formant changes of the repetitions were analysed (N=101), whereas the F2 locus equations are based on the original raw data (N=2040), therefore, one might argue that subtle effects might have been eliminated. This is an unlikely explanation considering that ΔF2 not even approached significance, as opposed to Whiteside and Varley's previous findings [8]. It seems more plausible to interpret F2 locus equations as preserving the acoustic context, i.e. the onset and midpoint frequencies, whereas ΔF2 represents only the change and strips said context away (see section 2.2). As movements are context dependent, both start and endpoint have to be considered; e.g. a rapid change in hand position of 30 cm has different contextual implications if the endpoint is a table top or the other hand. Therefore, we would like to argue that it is the context preserving F2 locus equations that represent speech motor control phenomena better than absolute formant changes.

Within the **F2 locus equations** both slope and R² were found to be functions of syllable frequency. The current data set underpins the suggestion that high frequency syllables are produced with greater degrees of coarticulation as shown by steeper **slope** values (see Figures 2 & 3 and Table 3). The analysis of **R²**, or goodness of fit, showed that high frequency syllables are produced with greater consistency than low frequency syllables.

Sex differences were neither found in the durational measures nor in the F2 locus equations [11]. This suggests that the effect of syllable frequency is similar for both male and female speakers. However, there were significant differences in the **absolute formant changes** in F1 and F2; both of which where higher for female speakers. This is either due to sociophonetic factors, as reported previously in the literature for formal settings and experimental conditions [15], or a physiological effect as suggested by Simpson (2001), who, based on a combination of acoustic and articulatory measures, shows sex differences in the *temporal* and *frequency domain* for a variety of American English [10]. The former is displayed in longer vowel durations for female speakers (by 8.9%); the latter in the absolute formant changes: for female speakers ΔF1 and ΔF2 values are 27.9% and 26.8% higher respectively; as calculated using the time indexes t1 and t3 which represent the onset and temporal midpoint of the first vowel in the diphthong (see Simpson 2001, 2158). This is comparable to the differences reported here in section 3.3 (37.5% and 31.8% respectively). Simpson suggests that because similar temporal and spectral differences were reported for different linguistic tasks (read and spontaneous speech) and different socio-cultural backgrounds (German [16] and British English [17]), the observed sex differences are more likely to be caused by physiological differences than by sociophonetic variation. Our findings corroborate Simpson's suggestion that physiological differences account for differences in absolute formant changes; irrespective of the linguistic task performed. We showed similar sex differences irrespective of the syllable frequency; viz. the speaker sex and syllable frequency interaction was not significant.

5 Conclusion

The significant differences between high and low frequency syllables, which showed in durational and coarticulation measures, are additional supporting evidence for the two syllable categories being encoded differently.

The overall consistency and faster production of high frequency syllables may be indicative of a more holistic phonetic encoding of syllable gestalts for this category. This deserves further exploration.

Acknowledgements

The authors wish express their sincere gratitude to the participants who devoted ten weeks to the main study, spent six of these listening daily to a CD and were locked into a sound proof booth for a grand total of just over two hours. Thank you!

References

[1] Keller, E. (1987). "Motor and sensory processes of language", Hillsdale, NJ, Lawrence Erlbaum Associates.

[2] Schiller, N.O., Meyer, A.S., et al. (1996). "A comparison of lexeme and speech syllables in Dutch", *Journal of Quantitative Linguistics* 3, 8-28.

[3] Cholin, J., W.J.M Levelt, et al (2006). "Effects of syllable frequency in speech production", *Cognition* 99(2), 205-235.

[4] Crompton, A. (1982). "Syllables and segments in speech production", in *Slips of the tongue and language production*, A. Cutler (ed.) Berlin: Mouton. 109-62.

[5] Levelt, W.J.M. and L. Wheeldon (1994). "Do speakers have access to a mental syllabary?", *Cognition* 50, 239-69.

[6] Browman, C. P. and L. Goldstein (1991). "Representation and realtiy: Physical systems and phonological structure", *Haskins Laboratory Status Report on Speech Research*, SR-105/106, 83-92.

[7] Whiteside, S. and R. Varley (2001). "What is the underlying impairment in acquired apraxia of speech?", *Aphasiology* 15 (1), 39-84.

[8] Whiteside, S. and R. Varley (1998a). "Dual-route phonetic encoding: a synthesis of acoustic evidence from normal speech", *Workshop on the sound patterns of spontaneous speech*, Aix-en-Provence, France, European Speech Communication Association (ESCA).

[9] Whiteside, S. and R. Varley (1998b). "A reconceptualisation of apraxia of speech: a synthesis of evidence", *Cortex* 34, 221-31.

[10] Simpson, A. (2001). "Dynamic consequences of differences in male and female vocal tract dimensions", *Journal of the Acoustical Society of America* 109 (5), 2153-64.

[11] Sussman, H.M., H. McCaffrey, et al. (1991). "An investigation of locus equations as a source of relational invariance for stop place categorization", *Journal of the Acoustical Society of America* 90 (3), 1309-25.

[12] Baayen, R.H., Piepenbrock, R., & Rijn, H. van (1993). "The CELEX Lexical Database (Release 1) [CD-ROM]", Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

[13] Krull, D. (1987). "Second formant locus patterns as a measure of consonant-vowel coarticulation", *PERILUS* V, 43-61.

[14] Sussman, H.M., Fruchter, D., et al. (1998). "Linear correlates in the speech signal: The orderly output constraint", *Behavioural and Brain Sciences* 21 (2), 241-99.

[15] Byrd, D. (1994). "Relations of sex and dialect to reduction", *Speech Communication* 15, 39-54.

[16] Simpson, A. (1998). *Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonologischen Theoriebildung*, Arbeitsbereiche des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK) 33.

[17] Whiteside, S. (1996). "Temporal-based acoustic-phonetic patterns in read speech: Some evidence for speaker sex differences", *Journal of the International Phonetic Association* 26, 23-40.