Foreign accent recognition based on temporal information contained in lowpass-filtered speech

Marie-José Kolly[1], *Adrian Leemann*[1], *Volker Dellwo*[1]

[1]Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, Switzerland
{marie-jose.kolly, adrian.leemann}@pholab.uzh.ch, volker.dellwo@uzh.ch

Abstract

Can the foreign accent of a speaker be recognized based on suprasegmental temporal information? For a perception experiment we created stimuli based on German sentences read by six French and six English speakers. These foreign-accented sentences were manipulated by (1) applying a lowpass filter with a cutoff frequency of 300 Hz and (2) applying the same lowpass filter and monotonizing F0. In a between-subject 2AFC perception experiment we tested the accent recognition ability of 15 Swiss German listeners per signal manipulation condition. The results showed that speakers' native language could be recognized above chance in both conditions. However, listeners obtained significantly lower recognition scores in the monotonized condition. Furthermore, higher recognition scores were obtained for French-accented speech in the monotonized condition, a result that is discussed in light of research on speech rhythm. We further report an effect for *speaker* within each accent group. The results suggest that suprasegmental temporal information allows for foreign accent recognition to some degree.

Index Terms: foreign accent recognition, speaker origin, lowpass-filtered speech, temporal characteristics, speech rhythm

1. Introduction

"Judging by your accent, you must be French" – people readily engage in foreign accent recognition tasks when listening to second language speech. But how, i.e. based on which cues, do listeners make decisions on a speaker' native language (L1)? Second language (L2) speech differs from native speech in a number of characteristics, and some of these characteristics are perceptually salient to listeners. For example, /r/ in the English word *foreign* is typically pronounced as a uvular trill [ʀ] or fricative [ʁ] by French speakers and as an alveolar trill [r] by Italian speakers. Provided that an English listener has common knowledge of French, an [ʀ] in *foreign* – among other cues – may lead him/her to guess the speaker's L1 as being French. Research has shown the importance of segmental cues for foreign accent recognition [1, 2].

The importance of suprasegmental cues for foreign accent recognition has been investigated by a handful of studies, which focused on frequency domain information. [3], for example, found that the absence of segmental accent-cues still allows listeners to recognize speaker origin in L2 speech, based on cues to f0 variability (i.e., intonation) and segment durations. [4] also demonstrated the importance of cues to f0 variability for foreign accent recognition. [4] further found that listeners were no longer able to recognize foreign accents in lowpass-filtered speech below 350 Hz, which suggested that time domain cues alone are not sufficient for this type of task. However, the multiple choice listening task used in [4] allowed the response "I don't know", an option that was frequently chosen by listeners. An alternative forced choice (AFC) experiment design may have yielded different results. Moreover, evidence from the field of dialect recognition suggested that time domain characteristics allow for dialect recognition: In a 4AFC experiment, listeners were able to recognize 3/4 Swiss German dialects in lowpass-filtered speech below 250 Hz [5].

The contribution of suprasegmental time domain information to foreign accent recognition was shown in [6]: Listeners were able to recognize foreign accents based on primarily temporal cues contained in 1-bit requantized speech [6, 7], for which the bit-rate of the acoustic signal was reduced to 1-bit, and in 6-band noise vocoded speech [8], for which amplitude envelopes were extracted from 6 frequency bands and used to modulate white noise. The latter sounds like a harsh whisper [9]. However, in signal manipulation conditions where listeners had no access to cues from the frequency domain (e.g. in 3-band noise vocoded speech, or in monotonized *sasasa*-speech [10]), foreign accent recognition was no longer possible [6]. The outcome of this research suggested that either it was the interplay between time and frequency domain characteristics that enabled foreign accent recognition, or that time-domain-only signal conditions that occur in natural situations would possibly yield different results and enable foreign accent recognition. In fact, 3-band noise vocoded speech and *sasasa*-speech are extremely distorted speech signals: In 3-band noise vocoded speech, the source signal of speech is replaced with white noise. In our *sasasa*-speech, every voiced speech interval was replaced with the same [a]-sound and every unvoiced speech interval with the same [s]-sound. Such "speech"-signals are unlikely to occur in everyday situations, which was mirrored by listeners' feedback in [6].

The present contribution, a follow-up experiment on [6], explores foreign accent recognition based on time domain characteristics contained in lowpass-filtered speech. This type of signal may appear more natural for listeners, since lowpass-filtered speech occurs in everyday situations: When a conversation is heard through a closed door, for example, or through a thick wall [11]. In this kind of situation, a listener may try to guess the language, accent or identity of the speakers. These guesses are confirmed once the speakers open the door: Their language, accent or identity becomes apparent to the listener. Listeners are therefore assumed to be familiar with the correspondence between unfiltered and filtered speech (e.g. of a particular language, accent, or speaker).

We aimed at using stimuli that contain no information on speech segmental content in order to isolate suprasegmental temporal and rhythmic features. We therefore filtered speech with a cutoff frequency of 300 Hz. We did not use a higher cutoff, since we wanted to exclude cues to vowel qualities: F1-values of vowels below 300 Hz are rather unusual in French, English and German [12, 13, 14]. We did not use a lower

cutoff, since female mean f0-values often attain 250 Hz in read speech [12, 15] and we wanted to include cues to f0 variability in one of our signal manipulation conditions – henceforth *lowpass* condition. We used the same filter for our second signal manipulation condition, and additionally monotonized f0 – henceforth *lowpass.monotonized* condition.

2. Materials and methods

2.1. Subjects

In a between-subject design, we tested a total of 30 Swiss German listeners: 15 listeners were tested with the *lowpass* condition (6 male / 9 female) and 15 with the *lowpass.monotonized* condition (5 male / 10 female). Subjects were university students and aged between 18 and 31 (M=23, SD=3). None of the subjects reported significant problems with hearing or sight. Their school education in second language French and English was comparable: French is usually introduced as a second, English as a third language in Swiss German schools. Swiss German university students have studied French and English for approximately 11 and 6 years respectively. We assumed listeners to have a similar level of familiarity with French and English speakers of German respectively, since the listener group was homogenous in terms of age and educational level.

2.2. Material

Stimuli were created based on speech from 6 French and 6 English native speakers (3 males / 3 females each). French speakers' self-assessed proficiency in German was intermediate (B1 to B2), English speakers' proficiency ranged from beginner to intermediate (A1 to B2), cf. [16]. Speakers' foreign accent degree was rated on a 5-point scale (1=very strong accent, 5=no accent) by 10 Swiss German listeners in a previous experiment. Results revealed that the degree of accentedness did not differ between the French and the English speakers [6].

Speakers read a list of 18 sentences that contained 12–16 syllables each. They were recorded with a *Fostex FR-2LE* solid-state recorder and a *Sennheiser MKE 2p-c* clip-on microphone (48 kHz, 16 bit) in a quiet room at the University of Zurich or in their own homes. We selected different sets of 9 sentences per speaker such that the experiment contained 108 sentences. Every sentence appeared 6 times in the experiment: 3 times with a French and 3 times with an English accent (cf. [6]).

Lowpass-filtered stimuli were constructed using Praat [17]. Every sentence was lowpass-filtered with a cutoff frequency of 300 Hz and a smoothing-value of 50 Hz (width between pass and stop, cf. [17]). An example of natural and lowpass-filtered speech from our stimuli is shown in Figure 1. To create monotonized lowpass-filtered speech, we first removed octave jumps automatically [17] and then replaced the pitch points of every sentence with the mean pitch value of the sentence. We used this procedure since averaging all male and all female sentences to a specific f0 mean produced stimuli that sounded unnatural (as judged by informal listening tests). We ran t-tests to examine the effect of the factor *accent* on mean f0: We did not find significant differences in f0 means between the French and the English accent group, neither for the *lowpass* (t=-0.53, ns, df=106) nor for the *lowpass.monotonized* condition (t=-0.08, ns, df=106). The accent recognition scores reported in section 3 are thus assumed to be independent from speakers' mean f0s. Finally, every stimulus-sentence was scaled to an intensity of 75 dB.

Informal perception experiments showed that listeners could not retrieve frequency domain information other than f0 variability from our stimuli. We presented listeners with two filtered vowel sounds and two categories as options: They were asked to decide which category belonged to which sound. Listeners were not able to identify single vowels in our lowpass-filtered speech. We understand this as evidence that our stimuli did not contain sufficient frequency domain cues that may have enabled the identification of individual vowel segments. Since frequency domain cues to consonants lie higher than 300 Hz this also means that consonantal distinctions could not be performed based on spectral envelope characteristics of consonants. In summary it can be said that our lowpass-filtered speech predominantly contained cues to voicing characteristics, i.e. on- and offset of voice as well as – in the *lowpass* condition – to changes of f0 over time (i.e., intonation).

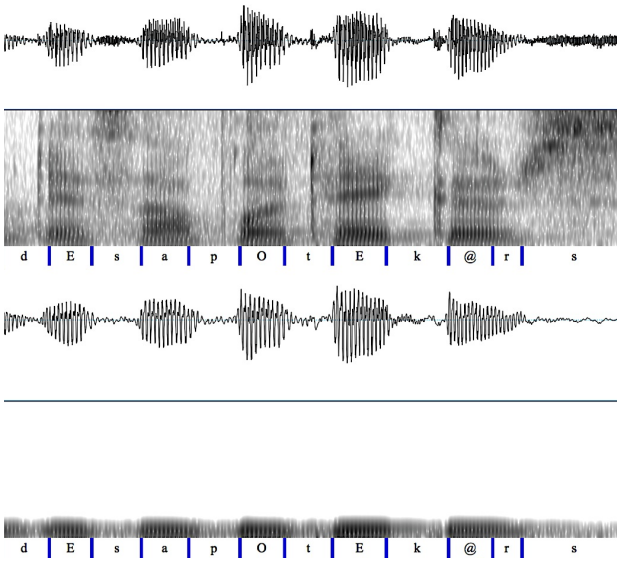

Figure 1: *SAMPA-transcribed waveform and spectrogram of the phrase* des Apothekers *'of the pharmacist' spoken by a native English speaker; natural (top) and lowpass-filtered (bottom) speech.*

2.3. Procedure

Listeners were tested in a quiet room at the University of Zurich. The experiment lasted between 15 and 25 mins. Subjects heard the stimuli through high-quality earphones, where the order of the stimuli was randomized separately for each subject. For both signal conditions, the sentence corresponding to the acoustic stimulus was presented on a laptop screen two seconds preceding the acoustic stimulus and during acoustic stimulus presentation. Following the presentation of each stimulus, subjects had to decide whether they heard French- or English-accented German by clicking on the corresponding button on a laptop computer, using the experiment interface shown in Figure 2. They further indicated the confidence of their response on a 3-point scale (1 = sure, 2 = rather unsure, 3 = only guessing).

Figure 2: *Experiment interface; to give their response, listeners clicked on one of the small blue rectangles.*

2.4. Data analysis and statistics

Based on each listener's responses we calculated *d'*, a measure derived from signal detection theory, based on the numbers of hits, false alarms, correct rejections and misses [18]. *d'* is obtained from each listeners' hit rate and false alarm rate: *d'* = *z-value*(hit)–*z-value*(false alarm). It measures listeners' sensitivity, i.e. their ability to discriminate two types of signals – French- vs. English-accented German – while canceling out response bias. Perfect sensitivity is reached at a *d'*-value of 4, whereas a *d'*-value of 0 indicates sensitivity at chance level. Normality of the *d'*-distribution was checked by visual inspection of quantile plots. To obtain listeners' recognition scores for each of the two signal types – i.e. accents – separately, we calculated the percentage of listeners' correct responses: *%correct* = (hits + correct rejections)/(hits + false alarms + correct rejections + misses). Statistical analyses were conducted using *R* [19]. We used two-sided t-tests and tested at a significance level of α=0.05.

3. Results

Results are presented as follows: In 3.1 we report the findings on listeners' general ability to recognize French- and English-accented German in *lowpass* and *lowpass.monotonized* speech. 3.2 shows the effect of signal manipulation condition on recognition performance. 3.3 presents results on listeners' recognition performance for French- and English-accented speech separately. 3.4 shows the effect of speaker on listeners' recognition performance.

3.1. Listeners' accent recognition performance

T-tests showed that listeners were able to recognize French- and English-accented German above chance in the *lowpass* condition (t=7.15, p<0.0001, df=14) as well as in the *lowpass.monotonized* condition (t=6.09, p<0.0001, df=14). This result is presented in Figure 3. Compared to *d'*-values of 4 for perfect sensitivity, the values reported here are fairly low (*lowpass*: M=0.61, *lowpass.monotonized*: M=0.39). However, this is in line with other investigations that use strongly degraded speech: [20], for example, report mean *d'*-values of 0.17 and 0.30 for listeners' recognition of English dialects in monotonized *sasasa*-speech.

3.2. Effect of signal manipulation condition

As can be seen in Figure 3, the two boxplots' interquartile ranges only overlap to a small degree: There was a significant difference between the signal conditions (t=2.04, p=0.05, df=26), with listeners obtaining higher accent recognition scores in *lowpass* (blue; M=0.61, SD=0.33) than in *lowpass.monotonized* (red; M=0.39, SD=0.25). Furthermore, the signal conditions differed significantly in listeners' certainty of response (t=-10.43, p<0.0001, df=3201), with listeners reporting higher degrees of certainty when making decisions about accents in *lowpass* (M=1.60, SD=0.68) as opposed to *lowpass.monotonized* (M=1.87, SD=0.76).

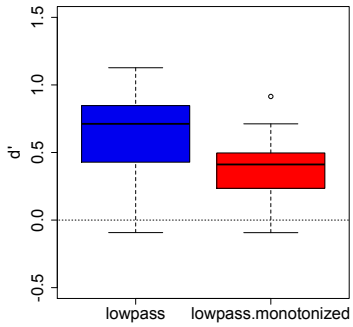

Figure 3: *Boxplots of* d' *for two signal conditions; dotted line = performance at chance.*

3.3. Effect of accent type

We calculated the percentage of correct responses for the French- and the English-accented stimuli separately: *%correct*. T-tests showed that French-accented German obtained higher recognition scores in *lowpass.monotonized* speech (red; t=-2.08, p<0.05, df=28) but not in *lowpass* speech (blue; t=-0.24, ns, df=28). This result is presented in Figure 4.

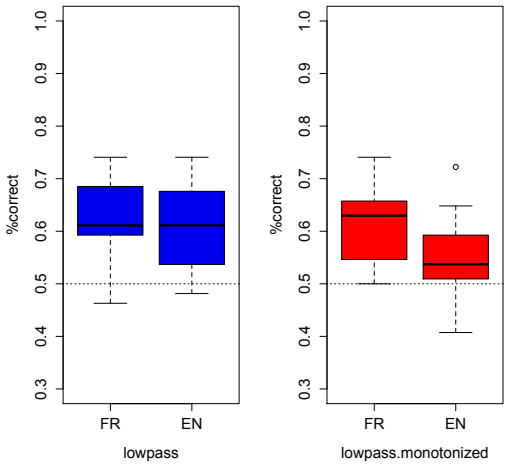

Figure 4: Boxplots of %correct for two accents by signal condition; dotted line = performance at chance.

3.4. Effect of speaker

A univariate ANOVA with *%correct* as the dependent variable shows that listeners' recognition scores differed depending on the speaker who articulated the sentences, within the French (F(5, 24)=9.04, p<0.01) as well as within the English (F(5, 24)=9.35, p<0.01) accent group (signal manipulation conditions pooled). This result is illustrated in Figure 5. We found a moderate correlation between *%correct* for each

speaker and speakers' accent degree (r=-0.43; French and English speakers pooled).

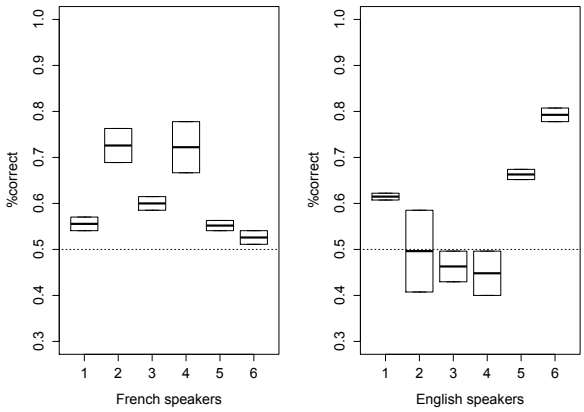

Figure 5: Boxplots of %correct for six French (left) and six English speakers (right); dotted line = performance at chance.

4. Discussion & Conclusion

Our results showed that listeners were able to recognize French- and English-accented speech above chance in the *lowpass* as well as in the *lowpass.monotonized* condition.

The result of the *lowpass* condition suggests that a 2AFC task allows listeners to recognize foreign accents when only cues to time domain and to f0 variability are available – which was not possible in [4], where listeners had the possibility to respond "I don't know". Our findings reflect similar results as research on dialect recognition in lowpass-filtered speech below 250 Hz [5], and findings on language discrimination by newborns in lowpass-filtered speech below 400 Hz [21] or by adults in lowpass-filtered speech below 180 Hz [22].

Our data on *lowpass.monotonized* speech shows that listeners are able to recognize foreign accents when no frequency domain information is present. Similar recognition performances were observed in 6-band noise-vocoded speech, a signal condition that allows listeners to access frequency domain information to some degree [6]. However, [6] showed that listeners performed at chance for signal manipulation conditions that did not contain frequency domain information, in particular for monotonized *sasasa*-speech based on voiced and voiceless intervals (see section 1), which contains similar suprasegmental temporal information as our *lowpass.monotonized* stimuli: Lowpass-filtered speech contains information about voice timing and information about intensity timing. *Sasasa*-speech contains information about voice timing only. Two explanations can be put forth for the discrepancy in listeners' accent recognition performance in these two signal conditions. (1) *Lowpass.monotonized* speech contains cues to intensity, which was not the case for the monotonized *sasasa*-speech used in [6]. The combination of time domain and intensity domain cues may have been important for listeners' ability to recognize foreign accents when no frequency domain information was available. (2) *Sasasa*-speech is unlikely to occur in natural situations; however, listeners can be assumed to be familiar with lowpass-filtered speech (cf. section 1), which may affect their recognition performance.

Our results further showed that listener performance and confidence differed significantly with regard to the signal manipulation condition: accent recognition performance as well as confidence was higher in *lowpass* than in *lowpass.monotonized* speech. It is plausible that this has to do with the fact that *lowpass* stimuli are signal-degraded to a lesser extent than *lowpass.monotonized* stimuli, i.e. they contain more cues – frequency domain cues in particular – that listeners can use to solve the accent recognition task. From this we infer that the absence of intonation in the *lowpass.monotonized* condition affected listener performance, but still allowed for accent recognition above chance. Similarly, [4] and [6] showed that listeners' accent recognition performance decreases as frequency domain information is reduced in signal-degraded speech.

We found that listeners' performance was significantly higher for the French-accented than for the English-accented stimuli in the *lowpass.monotonized*, but not in the *lowpass* condition. This suggests that French-accented German sounds perceptually more salient in the suprasegmental temporal domain than English-accented German. If interferences from speakers' L1 account for this, one may speculate that English is in fact closer to German than French in its suprasegmental temporal features – as it has been suggested by the literature on speech rhythm, which classified languages in rather "syllable-timed" (e.g. French) and rather "stress-timed" (e.g. English, German) [23–26], or in more and less "regular" [27].

We further found a significant effect of speaker on listeners' accent recognition performance, for the French- as well as for the English-accented stimuli. However, listeners' recognition performance for each speaker was only moderately correlated with speakers' accent degree. Since accent degree was rated based on natural speech (cf. [6]) it may be that listeners focused on different cues when listening to filtered speech, as reported in [28] – where it was found that listeners' ratings of foreign accent degree in natural speech and in filtered speech were not correlated. However, more research is needed before any conclusions can be drawn on our data.

Implications of this research can be found in the domain of second language acquisition: Our results suggest that suprasegmental temporal features are especially salient in French speakers' German speech. If an alleviation of foreign accentedness is desired, then learners of a second language that differs from their native language in its suprasegmental temporal organization may practice this type of feature in particular. From a more practical viewpoint, it has been shown that temporal features of foreign-accented speech have an effect on speakers' intelligibility [29].

Conducting research with more forensic phonetic applications in mind, we plan further perception experiments with our *lowpass* and *lowpass.monotonized* conditions without presenting visual information on sentence content. This task will be more similar to forensically relevant situations. For example, an ear-witness may hear a crime-related conversation through a closed door, and subsequently be asked to describe the linguistic profiles of the speakers s/he heard.

5. Acknowledgements

This research was supported by the Swiss National Science Foundation (SNSF; grant number: 100015_135287) and a grant of the Department of General Linguistics, University of Zurich. The authors would like to thank their subjects for their contribution to this experiment. We also thank Stephan Schmid for his expert advice on foreign-accented speech.

6. References

[1] Cunningham-Andersson, U. and Engstrand, O., "Perceived strength and identity of foreign accent in Swedish", Phonetica, 46:138-154, 1987.

[2] Boula de Mareüil, P., Vieru-Dimulescu, B., Woehrling, C. and Adda-Decker, M., "Accents étrangers et régionaux en français", Traitement Automatique des Langues, 49:135-163, 2008.

[3] Boula de Mareüil, P. and Vieru-Dimulescu, B., "The contribution of prosody to the perception of foreign accent", Phonetica, 63:247-267, 2006.

[4] Van Els, T. and De Bot, K., "The role of intonation in foreign accent", Modern Language Journal, 71:147-155, 1987.

[5] Leemann, A. and Siebenhaar, B., "Perception of dialectal prosody", Proceedings of Interspeech 2008, Brisbane, Australia: 524-527, 2008.

[6] Kolly, M.-J. and Dellwo, V., "Cues to linguistic origin: The contribution of speech temporal information to foreign accent recognition", Journal of Phonetics, 42:12-23, 2014.

[7] Licklider, J.C.R. and Pollack, I., "Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech", Journal of the Acoustical Society of America, 20:42-51, 1948.

[8] Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J. and Ekelid, M., "Speech recognition with primarily temporal cues", Science, 270:303-304, 1995.

[9] Scott, S.K., "The neurobiology of speech perception", in A. Cutler [Ed], Twenty-first century psycholinguistics: Four cornerstones, 141-156, Mahwah, NJ: Erlbaum, 2005.

[10] Ramus, F. and Mehler, J., "Language identification with suprasegmental cues: a study based on speech resynthesis", Journal of the Acoustical Society of America, 105:512-521, 1999.

[11] Hervais-Adelman, A.G., Davis, M.H., Johnsrude, I.S., Taylor, K.J. and Carylon, R.P., "Generalization of perceptual learning of vocoded speech", Journal of Experimental Psychology: Human Perception and Performance, 37:283-295, 2011.

[12] Peterson, G.E. and Barney, H.L., "Control methods used in a study of the vowels", Journal of the Acoustical Society of America, 24:175-184, 1952.

[13] Hillenbrand, J., Getty, L.A., Clark, M.J. and Wheeler, K., "Acoustic characteristics of Amercian English vowels", Journal of the Acoustical Society of America, 97:3099-3111, 1995.

[14] Delattre, P., Comparing the phonetic features of English, German, Spanish and French, Heidelberg: Julius Groos, 1965.

[15] Künzel, H.J., Masthoff, H.R. and Köster, J.P., "The relation between speech tempo, loudness, and fundamental frequency: an important issue in forensic speaker recognition", Science and Justice, 35:291-295, 1995.

[16] Council of Europe, Common European framework of reference for languages: learning, teaching, assessment, http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf, accessed on 12 Jun 2014.

[17] Boersma, P. and Weenink, D., Praat: doing phonetics by computer, www.praat.org, 2013.

[18] Green, D.M, and Swets, J.A., Signal detection theory and psychophysics, New York: Wiley, 1966.

[19] R Core Team, R: A language and environment for statistical computing, version 3.0.1, http://www.R-project.org, 2013.

[20] White, L., Mattys, S.L. and Wiget, L., "Language categorization by adults is based on sensitivity to durational cues, not rhythm class", Journal of Memory and Language, 66:665-679, 2012.

[21] Nazzi, T., Bertoncini, J. and Mehler, J., "Language discrimination by newborns: toward an understanding of the role of rhythm", Journal of Experimental Psychology: Human Perception and Performance, 24:756-766, 1998.

[22] den Os, E., Rhythm and tempo of Dutch and Italian, Utrecht: Elinkwijk, 1988.

[23] Abercrombie, D., Elements of general phonetics, Edinburgh: Edinburth University Press, 1967.

[24] Pike, K., The intonation of American English, Ann Arbor: University of Michigan Press, 1945.

[25] Ramus, F., Nespor, M. and Mehler, J., "Correlates of linguistic rhythm in the speech signal", Cognition, 73:265-292, 1999.

[26] Grabe, E. and Low, E.L., "Durational variability in speech and the Rhythm Class Hypothesis", in C. Gussenhoven and N. Warner [Eds], Laboratory Phonology 7, 515-545, Berlin/New York: Mouton de Gruyter, 2002.

[27] Dellwo, V., "The role of speech rate in perceiving speech rhythm", Proceedings of Speech Prosody 2008, Campinas, Brazil: 275-278, 2008.

[28] Munro, M., "Nonsegmental factors in foreign accent: ratings of filtered speech", Studies in Second Language Acquisition, 17:17-34, 1995.

[29] Tajima, K., Port, R. and Dalby, J., "Effects of temporal correction on intelligibility of foreign-accented English", Journal of Phonetics, 25:1-24, 1997.