Listeners use temporal information to identify French- and English-accented speech

Marie-José KOLLY[a,b]*, Philippe BOULA DE MAREÜIL[b], Adrian LEEMANN[c], Volker DELLWO[a]

*Corresponding author

marie-jose.kolly@uzh.ch
philippe.boula.de.mareuil@limsi.fr
al764@cam.ac.uk
volker.dellwo@uzh.ch

[a] Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, Plattenstrasse 54, 8032 Zurich, Switzerland

[b] Laboratoire d’Informatique pour la Mécanique et les Sciences de l’Ingénieur (LIMSI), CNRS, Université Paris-Saclay, Rue John von Neumann, 91405 Orsay Cedex, France

[c] Phonetics Laboratory, Department of Theoretical and Applied Linguistics, University of Cambridge, Sidgwick Avenue, Cambridge, CB3 9DA, United Kingdom

Abstract

Which acoustic cues can be used by listeners to identify speakers’ linguistic origins in foreign-accented speech? We investigated accent identification performance in signal-manipulated speech, where (a) Swiss German listeners heard native German speech to which we transplanted segment durations of French-accented German and English-accented German, and (b) Swiss German listeners heard 6-band noise-vocoded French-accented and English-accented German speech to which we transplanted native German segment durations. Therefore, the foreign accent cues in the stimuli consisted of only temporal information (in a) and only strongly degraded spectral information (in b). Findings suggest that listeners were able to identify the linguistic origin of French and English speakers in their foreign-accented German speech based on temporal features alone, as well as based on strongly degraded spectral features alone. When comparing these results to previous research, we found an additive trend of temporal and spectral cues: identification performance tended to be higher when both cues were present in the signal. Acoustic measures of temporal variability could not easily explain the perceptual results. However, listeners were drawn towards some of the native German segmental cues in condition (a), which biased responses towards ‘French’ when stimuli featured uvular /ʀ/s and towards ‘English’ when they contained vocalized /ʀ/s or lacked /ʀ/.

Keywords

accent identification, speech rhythm, temporal cues, segmental cues, additivity of cues, forensic phonetics

1 Introduction

“Judging by your accent, you must be French” – people regularly engage in foreign accent identification tasks in everyday social interactions. Which acoustic cues are useful for such tasks? The question is particularly relevant when the origin of an individual has to be determined for legal cases, where forensic phoneticians or ear-witnesses establish a speaker’s profile to reduce the number of potential suspects (Ellis, 1994; Köster et al., 2012). Aside from forensic caseworkers, a number of governmental institutions conduct Linguistic Analyses for the Determination of the geographical Origin (LADO) of an individual. Here, an asylum seeker’s claim to originate from a particular region is examined, when no valid identification documents are available (Baltisberger and Hubbuch, 2010). Foreign accent identification can be a crucial part of speaker profiling and LADO, as some individuals use second language speech to disguise their native language and thus their geographical origin (Cambier-Langeveld, 2010).

Foreign-accented speech contains a large number of specific features, and some of these are perceptually salient in terms of geographical origin. The most salient features indicative of a foreign accent are likely to be found on the segmental level (Boula de Mareüil et al., 2004a; Boula de Mareüil et al., 2008; Cunningham-Andersson and Engstrand, 1989; Flege and Port, 1981; Vieru et al., 2011). /R/ in the Swiss German toponym *Zürich*, for example, is typically realized as a uvular trill [ʀ] or fricative [ʁ] by French speakers, and as an alveolar approximant [ɹ] by English speakers – as opposed to the Zurich Swiss German articulation of an alveolar trill [R] or tap [ɾ] (Werlen, 1980). Foreign-accented speech is characterized, to some extent, by interferences from the speakers’ first language. Based on such interferences, for example in the /R/ realization, listeners can typically guess the native language (i.e., French, English, Swiss German) of the speaker.

In some adverse listening situations, access to segmental cues is reduced. One can think of speech that was recorded through a closed door, on a mobile telephone, or in a noisy environment, as typically encountered in the domain of forensic phonetics: telephone speech is involved in 90% of forensic phonetic casework (Hirson et al., 1995), and speech material for LADO, too, is often obtained over a landline network (Baltisberger and Hubbuch, 2010). Forensic caseworkers’ decisions must most often rely on degraded segmental cues and/or on other cues. Here, speech prosodic information might play a crucial role: Listeners’ ability to recognize words, for example, was shown to strongly deteriorate in noise, while their ability to recognize prosodic patterns remained unaffected by it (Van Zyl and Hanekom, 2011). However, adverse listening conditions often also reduce certain types of prosodic features, particularly features from the frequency domain. When speech is transmitted through a mobile telephone, for example, the frequency range is reduced to a frequency band between 350 and 3200 Hz (Künzel, 2001), measurements of vowel qualities are obscured (Byrne and Foulkes, 2004), and speakers’ fundamental frequency tends to be higher due to speaking more loudly on the telephone (ibid.). *Temporal* cues are typically less affected by distortions of the speech signal as they occur in telephone speech (Chen et al., 2005; Leemann et al., 2014). In the context of the present paper, we use the term *temporal* to refer to durations of speech segments, as this is the feature that we manipulated in our stimuli. Segment durations have an effect not only on segmental but also on suprasegmental timing patterns (van Santen and Shih, 2000).

Can listeners identify the origin of speakers based on temporal features of their non-native speech? A rationale for this idea comes from the domain of speech rhythm research – the study of the suprasegmental temporal organization of speech. Languages have been argued to differ in their rhythm (Abercrombie, 1967; Lloyd James, 1929; Pike, 1945). The acoustic features that allegedly correlate with the perception of speech rhythm remain to be fully determined, as rhythm metrics proposed in the literature were reported to be influenced not only by language (Dellwo, 2006; Grabe and Low, 2002; Ramus et al., 1999) or dialect (Ferragne and Pellegrino, 2004; Leemann et al., 2012; White and Mattys, 2007b), but also by factors such as speaker, sentence material, or annotator (Arvaniti, 2012; Dellwo et al., 2015; Leemann et al., 2014; Vieru et al., 2011; Wiget et al., 2010). Numerous studies reported that listeners are sensitive to suprasegmental temporal information contained in speech (e.g. Pinet and Iverson, 2010; Quené and van Delft, 2010; Tajima et al., 1997). Furthermore, listeners were reported to use such information to distinguish between languages (Nazzi et al., 1998; Ramus and Mehler, 1999; Ramus et al., 2003) or dialects (adults: White et al., 2012; infants: White et al, 2014). It is thus conceivable that suprasegmental temporal information might be a potential cue to foreign accents such as French-accented and English-accented German.

French and English differ in their suprasegmental temporal organization. For example, English features higher durational variability between prominent and less prominent syllables than French (Delattre, 1966; Fant et al., 1991). French and English also differ on the segmental temporal level: English, but not French, features distinctive vowel quantity and vowel reduction; English has more complex syllables and consonant clusters than French (Auer, 2001; Dauer, 1983; German shows similar temporal features as English in these examples). Speakers of both French and English produce longer vowels before voiced than before unvoiced consonants, but this effect is stronger for English speakers (Laeufer, 1992). These segmental temporal differences between the two languages may translate to differences in suprasegmental temporal structure as well (van Santen and Shih, 2000). For example, listeners were shown to perceive French as more regularly timed than English or German (Dellwo, 2008). Furthermore, some of the temporal patterns discussed are typically carried over to a non-native language (Arslan and Hansen, 1997; McAllister et al., 2002). Voice Onset Time (VOT), for instance, is known to differ between French and English, and Hazan and Boulakia (1993) reported that bilingual speakers of French and English often produce VOT according to their dominant language. In conclusion, we start from the assumption that French-accented German and English-accented German differ in their segmental and suprasegmental temporal organization. We therefore hypothesize that listeners may be able to use such temporal features to identify the two accents.

The question whether particular foreign accents can be identified based on temporal cues has been studied only to a minor extent. Previous research on foreign accent identification more often than not featured material that contained a certain amount of frequency domain information in addition to temporal information: segment durations and intonation in prosody-transplanted speech (Boula de Mareüil and Vieru-Dimulescu, 2006); segment durations and degraded spectral features in 1-bit requantized speech (Kolly and Dellwo, 2014); temporal features of the amplitude envelope and degraded spectral features in 6-band noise-vocoded speech (Kolly and Dellwo, 2014); and temporal features of the amplitude envelope and of voicing in monotonized lowpass-filtered speech below 300 Hz (where some spectral features below 300 Hz may have been useful for accent identification; Kolly et al., 2014). In this line of research, listeners were reported to respond at chance level when stimuli contained (almost) no spectral features, e.g. in 3-band noise-vocoded speech and in monotonized *sasasa*-speech (see below; Kolly and Dellwo, 2014). The signal conditions discussed preserve mainly temporal features and different degrees of rudimentary spectral information. Findings showed that accent identification performance decreased with higher degradation of spectral features. The outcome of this research can be interpreted in two ways: on the one hand, the additivity of cues may have played a role, where the combination of temporal and spectral features potentially boosted identification performance (Du et al., 2011; Hjalmarsson, 2011). Listeners might, for example, identify an accent because some rudimentary spectral information occurs at a specific (and expected) moment in time. If the temporal integrity of the signal were completely degraded, the same spectral information might be of less or no use to the listener. Similarly, if the spectral information were completely absent, the temporal information, still intact, may be of less or no use to a listener (Dellwo, 2010). On the other hand, temporal information alone might allow for foreign accent identification if it were presented in a signal condition that occurs in natural listening situations. In fact, 3-band noise-vocoded speech and *sasasa*-speech are highly distorted signals: The process of noise-vocoding replaces the source signal of speech with white noise (Shannon et al., 1995), and, in the *sasasa*-experiment, every voiced interval was replaced with the same [A]-sound and every unvoiced interval with the same [S]-sound. ‘Speech’-signals such as these do not occur in everyday listening situations. It thus seems plausible that, because of a lack of experience with such signals, listeners are not able to interpret the temporal information contained in them.

To test whether listeners rely on the additivity of temporal and spectral cues to identify foreign accents, we separated both cues contained in the 6-band noise-vocoded speech used by Kolly and Dellwo (2014). We conducted two perception experiments to investigate if listeners can identify foreign accents (a) based on temporal features alone (henceforth *timeOnly*), and (b) based on strongly degraded spectral features alone (henceforth *freqOnly*). To isolate temporal features for (a), and to eliminate temporal features for (b), we used a signal manipulation frequently referred to as 'prosody transplantation'. The method was introduced by Osberger and Lewitt (1979) and has mostly been applied to investigate the importance of temporal and/or fundamental frequency patterns for the intelligibility of deaf speakers (Maassen and Povel, 1985; Osberger and Lewitt, 1979) and the intelligibility and/or degree of accentedness in non-native speech (Holm, 2008; Pinet and Iverson, 2010; Quené and van Delft, 2010; Rognoni and Busà, 2014; Tajima et al., 1997; Vitale et al., 2014; Winters and O'Brien, 2013). Prosody-transplanted speech has also been used to investigate whether segmental or prosodic cues are more important to identify foreign accents; findings suggest that segmentals prevail in the identification of native vs. Arabic- or Kabyle-accented French (Boula de Mareüil et al., 2004a), whereas prosody plays more into the identification of Spanish-accented Italian vs. Italian-accented Spanish (Boula de Mareüil et al., 2004b; Boula de Mareüil and Vieru-Dimulescu, 2006).

For the signal condition *timeOnly*, we transplanted segment durations of French- and English-accented German to native German, i.e., we modified German segment durations to match the segment durations of French- and English-accented German. This eliminated all spectral features of the foreign accents, while keeping the resulting stimuli fairly natural-sounding. For the signal condition *freqOnly*, we transplanted native German segment durations to French- and English-accented German, which eliminated all segmental and suprasegmental temporal information of the foreign accents. We then 6-band noise-vocoded the material in such a way that it contained the spectral information from 6-band noise-vocoded speech (Kolly and Dellwo, 2014). Apart from the fact that it allowed us to test effects of cue additivity, 6-band noise-vocoding was also performed to reduce spectral information, as it seemed plausible that intact spectral cues alone would lead to near-ceiling effects in perception experiments. A drawback of using the prosody transplantation and noise-vocoding approach is the artificiality of stimuli: the noise-vocoded speech of the *freqOnly* stimuli sounds highly unnatural; *timeOnly* speech sounds relatively natural but combines native frequency domain features with non-native temporal features, a hybrid signal that listeners also do not encounter in natural environments. However, this seems to be the ecologically most valid way of separating temporal and spectral features.

Our approach was (a) to test, in a perception experiment, whether listeners can recognize French- and English-accented German based on temporal features or spectral features of the foreign accents only, and (b) to investigate acoustic correlates that may explain listeners' behavior. In perception experiments, Swiss German listeners heard French- and English-accented *timeOnly* or *freqOnly* sentences and had to decide whether they heard a French or an English accent. We used a between-subjects design in which each signal condition was tested with different listeners, given that listeners may adapt to manipulated speech: Davis et al. (2005), for example, reported that the intelligibility of noise-vocoded speech increased with training. In the context of the present study, a within-subjects design may have encouraged listeners to use their familiarization with the sentence, speaker and accent characteristics from, say, the *timeOnly* experiment when completing the task in the *freqOnly* experiment, resulting in artifacts, as such information would have been of no use to them. To allow for a comparison with previous experiments, we used the recordings and experiment design from Kolly and Dellwo (2014). A number of acoustic temporal measures were applied to unmanipulated speech and to our stimuli in order to verify that duration transplantation had the desired effect on the the material. Furthermore, these acoustic temporal measures were used to explore potential acoustic correlates of listeners' accent identification performance.

2 Materials and methods

2.1 Subjects

A total of 40 native Swiss German listeners (16 male, 24 female) took part in the accent identification experiments. Listeners were University of Zurich students aged between 18 and 45 years (M=23.30, SD=4.37). None of them reported hearing disorders or problems with sight. Due to listeners' age, origin and educational level, we assumed a comparable level of familiarity with French and English speakers of German. Likewise, we presupposed similar levels of proficiency in French and English, as French is usually introduced as a second and English as a third language in Swiss German schools: Subjects had studied French and English for about 11 and 6 years, respectively. Before starting university studies, Swiss German students such as our subjects pass an exam called *Maturität* (*Baccalaureate*), for which their proficiency in French and English is expected to correspond to B2–C1 according to the Common European Framework for Languages (Erziehungsdirektion des Kantons Bern, 2009; Council of Europe, 2013). At university, students tend to use English more than French.

For our between-subject design, listeners were randomly attributed to two groups. We tested 20 listeners (10 male, 10 female) with the signal condition *timeOnly* and 20 listeners (6 male, 14 female) with the signal condition *freqOnly*.

2.2 Materials

2.2.1 Speakers

We collected Standard German speech from 18 speakers: three male and three female speakers for each language (French, English and Zurich German). Speakers' age ranged between 23 and 56 years (M=30.78, SD=8.02). The Zurich German speakers grew up in the city of Zurich; the French speakers in the French-speaking part of Switzerland; the English speakers in the US or in Canada, one female speaker in the UK (their English varieties feature similar durational patterns, for instance vowel reduction; Grenon and White, 2008; Shearme and Holmes, 1961; Tiffany, 1959).

Native Standard German speech for duration transplantation was obtained from Zurich German speakers, as our listeners were mostly Zurich German, too. In diglossic German-speaking Switzerland, dialects are used mainly for verbal communication, whereas Standard German is mainly used in the written form and in more formal oral situations (Ferguson, 1959; Kolde, 1981). The pronunciation of /R/ in Swiss Standard German is variable (Hove, 2002): some speakers produce an alveolar [R] or [ɾ], the variant present in most of the Swiss German dialects (Werlen, 1980); others produce /R/ as a uvular trill [ʀ] or fricative [ʁ]. In specific phonotactic positions, certain speakers may vocalize /R/ to schwa [ɐ], particularly in post-vocalic contexts, which corresponds to the Standard German system (Kohler, 1990). The Zurich German speakers recorded for the present experiments all used uvular as well as vocalized /R/ variants in their Standard German.

The Zurich German speakers used Standard German on a regular basis. French and English speakers self-assessed their proficiency in German using the Common European Framework for Languages (Council of Europe, 2013). French speakers' proficiency ranged between B1 and B2, English speakers' between A1 and B2. The origin and strength of their foreign accent was rated by 16 listeners (9 male, 7 female) in natural speech, on a 5-point scale (1=very strong accent, 2=strong accent, 3=medium accent, 4=slight accent, 5=no accent). Listeners' age ranged between 20 and 36 years (M=26.25, SD=5.20). None of the listeners was part of the group of subjects presented in Section 2.1. We constructed a linear model of *accent strength* as a function of *accent* and found no significant differences in *accent strength* between the French and the English speaker group (LM: F(1,10)=0.39, p=0.55; French speakers: M=2.86, SD=0.19; English speakers: M=2.67, SD=0.71; cf. Section 2.6 for details on statistical analyses). We further found their foreign accents to be recognized with high performance, in natural speech, as measured by *A'* (M=0.95, SD=0.03). This illustrates that speakers provided typical examples of French- and English-accented speech, and corresponds to the judgement of expert phoneticians (authors).

2.2.2 Reading materials and recordings

All speakers read a list of 18 Standard German sentences, which varied between 12 and 16 syllables (cf. Appendix). Prior to the recording, speakers familiarised themselves with the materials by reading the sentences aloud. The French and English speakers were recorded in a quiet room using a *Fostex FR-2LE* solid-state recorder (48 kHz; 16 bit) and a *Sennheiser MKE 2p-c* clip-on microphone. The Zurich German speakers were recorded in a sound-treated booth using a *Neumann STH-100* transducer microphone (44.1kHz; 16 bit). We selected a different set of 9 sentences from each French and English speaker to avoid identical sentence sets for all speakers and thus to obtain more variability of linguistic material in the experiment (Kolly and Dellwo, 2014). The experiment contained 108 sentences in total (2 accents × 6 speakers × 9 sentences): Each of the 18 Standard German sentences appeared six times in the experiment, three times read by a French speaker and three times read by an English speaker.

2.2.3 Segmentation

The 108 non-native sentences and their native German counterparts were segmented, on a phonetic level, by a trained phonetician (first author), using Praat (Boersma and Weenink, 2014). Segmentation and labelling decisions were based on visual inspection of waveforms and spectrograms, and on auditory criteria. All interval boundaries were placed at positive zero-crossings. In order to obtain an optimal transplantation of durational patterns, diphthongs and affricates were segmented into their components, glottal stops or laryngealized parts were treated as individual segments, and silent pauses were annotated without the application of a particular duration threshold. However, stops were not divided in separate closure and release sections.

2.3 Stimuli

We chose to transplant segment durations rather than syllable durations (e.g. Maassen and Povel, 1985; Osberger and Lewitt, 1979; Winters and O'Brien, 2013) since French- and English-accented German may differ on a very detailed durational level (cf. Section 1). Furthermore, segmental durations have been suggested to be an important cue for foreign accent identification (Kolly and Dellwo, 2014). Segment durations of the speech material read by each particular French and English speaker were therefore transplanted to material read by a native speaker (*timeOnly*) and vice versa (*freqOnly*).

Since we transplanted durational features, speaker pairs (French-Zurich German; English-Zurich German) were built according to a gender-specific ranking of articulation rate (cf. Table 1), as measured by *ratePeak* (cf. Section 2.4). In doing so, we avoided an extreme stretching of segments – which may result in artifacts such as chirp or whistle sounds – wherever possible (Quené and van Delft, 2010). For the signal condition *timeOnly*, for example, segment durations of FR04 (and those of EN01) were transplanted to ZH07.

Gender	**French speakers**	**Zurich German speakers**	**English speakers**
male	FR04 (M=5.41, SD=0.85)	ZH07 (M=5.55, SD=0.42)	EN01 (M=5.50, SD=0.43)
	FR01 (M=4.82, SD=0.71)	ZH14 (M=5.21, SD=0.54)	EN06 (M=5.26, SD=0.81)
	FR10 (M=4.14, SD=0.30)	ZH15 (M=5.11, SD=0.47)	EN07 (M=5.14, SD=0.54)
female	FR05 (M=5.09, SD=0.35)	ZH69 (M=5.63, SD=0.39)	EN03 (M=4.98, SD=0.45)
	FR03 (M=5.08, SD=0.58)	ZH71 (M=5.37, SD=0.52)	EN02 (M=4.83, SD=0.59)
	FR08 (M=4.75, SD=0.43)	ZH70 (M=5.22, SD=0.58)	EN04 (M=4.17, SD=0.63)

Table 1: Ranking of male and female speakers according to articulation rate as measured by *ratePeak*.

After the segmentation (cf. Section 2.2.3) we checked whether the matching versions of each sentence from each speaker pair (e.g. speakers FR04 and ZH07, sentence 03) were segmented into the same number of intervals, a prerequisite for the transplantation of segment durations. The number of intervals differed between the versions if either the number of segments, or the number of silent pauses was different. If only one version of a sentence featured a silent pause at a specific position, we introduced a silent part of the same length and at the same position to the other version, and added an interval to its segmentation. The silent part that was introduced was taken from the (silent) start or end of the sentence into which it was introduced, in order to obtain a maximally natural auditory effect (Pettorino and Vitale, 2012). In cases where the segment count was different, we merged intervals in the version that contained a higher number of segments, which resulted in some intervals containing multiple segments (cf. Figure 1). Intervals were merged according to syllable or phoneme boundaries, such that durational features of a syllable or phoneme would be transplanted to the same syllable or phoneme of the matching version of the sentence (Tajima et al., 1997). Typical examples for situations where intervals were merged are the following:

- Elisions:
    - Some native German speakers elided the schwa before a sonorant in unstressed syllables (e.g. *Regen* [ˈʁEː gN] vs. [ˈʁEː gəN] ‘rain’). In such cases, the schwa and the following sonorant of the French or English speaker’s sentence were merged into a single interval, as exemplified in Figure 1.
    - Some French or English speakers elided linking elements between the two components of a German compound (e.g. *Zahlungsbilanz* [ˈTSAː LʊŋBIˌLANTS] vs. [ˈTSAː LʊŋSBIˌLANTS] ‘balance of payment’). In such cases, the linking element and the preceding phone of the Zurich German speaker’s sentence were merged into a single interval.
- Epentheses: Some French and English speakers produced a velar plosive after velar nasals (e.g. *lange* [ˈLAŋgə] vs. [ˈLAŋə] ‘long’). In such cases, the nasal and the following plosive were merged into a single interval.

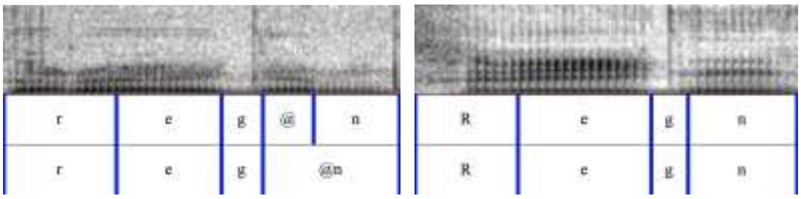

Figure 1: Annotation of segments (tier 1) and parallel annotation of two matching versions of a sentence resulting in merged segments (tier 2) for an English-accented (left spectrogram and annotation) and a native German (right spectrogram and annotation) token of *Regen* ‘rain’.

Figure 2 illustrates the signal processing steps undertaken to obtain stimuli for *timeOnly*: Segment durations of the native version of a sentence were modified with segment durations of its non-native counterpart. Native German speech intervals were therefore either stretched or compressed by means of Pitch Synchronous Overlap and Add (PSOLA) resynthesis, using a Praat script adapted from Boula de Mareüil and Vieru-Dimulescu (2006). The speech signal, albeit carrying some artifacts due to the stretching of particular segments, is still intelligible and rather natural. To obtain stimuli for *freqOnly*, segment durations of the non-native version of a sentence were modified with segment durations of its native counterpart. Sentences were subsequently 6-band noise-vocoded. We divided the speech signal into six logarithmically-spaced frequency bands. We used the same respective cutoff frequencies to filter white noise. The amplitude envelope was extracted from each speech band and multiplied with the corresponding noise band. The six noise bands were summed up to obtain 6-band noise-vocoded speech (cf. Kolly and Dellwo, 2014, for more detail). All stimuli were scaled to an intensity of 70 dB.

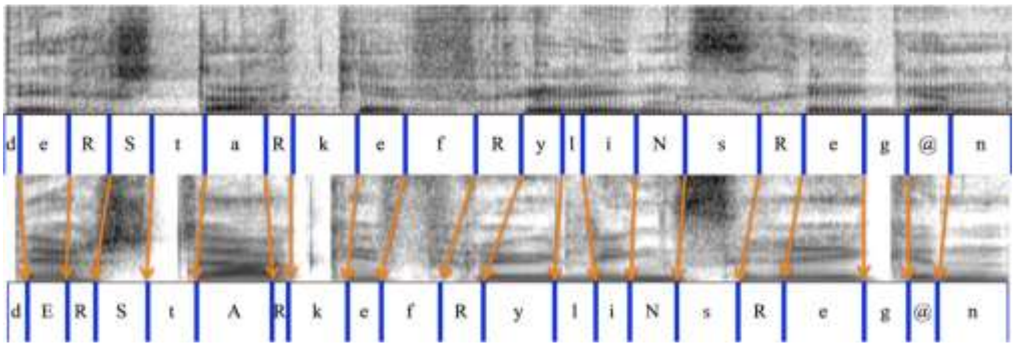

Figure 2: Modification of native German segment durations (bottom spectrogram and annotation) with French segment durations (top spectrogram and annotation). The phrase reads *der starke Frühlingsregen* 'the strong spring rain'.

2.4 Temporal measures applied

In the following, we present a number of acoustic measures that were applied to the material (i) to describe our stimuli and therefore verify what effect the duration transplantation may have had on certain durational characteristics of the material (cf. Section 3.2.1) and (ii) to explore potential acoustic correlates of listeners' identification performance (cf. Section 3.2.2). For this, we applied five different types of temporal measures to the natural and the duration transplanted speech: (1) measures of articulation rate, (2) pausing measures, and (3) a number of rhythm metrics based on the durational variability (3a) of vocalic and consonantal intervals, (3b) of voiced and voiceless intervals and (3c) of intervals between peaks in the amplitude envelope.

(1) Measures of articulation rate:

- *rateCV*, the number of consonantal and vocalic intervals per second (Dellwo, 2008);
- *ratePeak*, the number of automatically detected peaks in the amplitude envelope (Dellwo et al., 2012; Mermelstein, 1975), which roughly corresponds to the number of syllables, per second.

(2) Measures of pausing (Bosker et al., 2014; Cucchiarini et al., 2002; de Jong et al., 2013; Künzel, 1997):

- *pauseNbr*, the number of silent pauses;
- *pauseDur*, silent pause durations.

(3a) Rhythm metrics based on durational features of vocalic and consonantal intervals (derived from segmentation):

- *%V*, the percentage of time over which speech is vocalic (Ramus et al., 1999);
- *varcoVln*, the rate-normalized standard deviation of vocalic interval durations (*varcoV*: White and Mattys, 2007a), calculated on log-transformed interval durations;
- *nPVI_V*, the rate-normalized average difference between consecutive vocalic interval durations (Grabe and Low, 2002);
- *varcoC*, the rate-normalized standard deviation of consonantal interval durations (Dellwo, 2006);
- *nPVI_C*, the rate-normalized average difference between consecutive consonantal interval durations (Grabe and Low, 2002).

(3b) Rhythm metrics based on durational features of voiced and unvoiced intervals (automatically calculated using the default pitch detection algorithm in Praat):

- *%VO*, the percentage of time over which speech is voiced (Dellwo et al., 2007);
- *varcoVOln*, the rate-normalized standard deviation of voiced interval durations (*varcoVO*: Dellwo et al., 2007), calculated on log-transformed interval durations;
- *nPVI_VO*, the rate-normalized average difference between consecutive voiced interval durations (Dellwo et al., 2007);
- *varcoUV*, the rate-normalized standard deviation of unvoiced interval durations (Dellwo et al., 2007);
- *nPVI_UV*, the rate-normalized average difference between consecutive unvoiced interval durations (Dellwo et al., 2007).

(3c) Rhythm metrics based on durational features of intervals between automatically detected peaks in the amplitude envelope (one peak per vocalic segment):

- *varcoPeak*, the rate-normalized standard deviation of interval durations between automatically extracted amplitude peaks (Dellwo et al., 2012);
- *nPVI_Peak*, the rate-normalized average difference between consecutive interval durations between automatically extracted amplitude peaks (Dellwo et al., 2012).

The measures *varcoV* (White and Mattys, 2007a) and *varcoVO* (Dellwo et al., 2007) were calculated based on log-transformed interval durations, since the distributions of vocalic and voiced intervals were strongly positively skewed. Temporal measures were calculated sentence-by-sentence using the Praat plugin *Duration Analyzer* (available at http://www.pholab.uzh.ch/static/volker/software/plugin_durationAnalyzer.zip).

2.5 Procedure

Listeners were tested in a quiet room at the University of Zurich using a laptop computer. They heard stimuli over high-quality closed *Beyerdynamics DT 770 PRO* headphones, and stimulus order was randomized for each listener. Listeners tested for *freqOnly* heard strongly distorted speech; they were thus presented with sentence transcripts corresponding to each acoustic stimulus, which allowed them to parse the acoustic information (Davis et al., 2005). Sentence transcripts were presented on the computer screen two seconds prior to the acoustic stimulus, and remained on the screen during stimulus presentation. Listeners tested in the *timeOnly* signal condition were not given sentence transcripts, as the stimuli presented were readily intelligible. We cannot exclude that the display of sentence transcripts distracted listeners' attention from the acoustic signal in *freqOnly*; listeners tested with *timeOnly*, on the other hand, could focus their entire attention on the acoustic stimulus. Prior findings by Kolly and Dellwo (2014) suggest, however, that this potentially distracting effect is small compared to the gain from listeners being aware of the sentence content: Listeners identified accents above chance in 6-band noise-vocoded speech when the acoustic stimuli were presented with sentence transcripts, but not when they were missing.

Listeners were instructed as follows: they would hear Standard German sentences spoken by French and English speakers and they would have to decide, for each sentence, whether they heard French- or English-accented German, and how confident they were concerning their response. They were encouraged to respond intuitively. Listeners tested in the *freqOnly* signal condition were additionally informed that they would hear manipulated speech and that they would be able to read the sentence corresponding to the acoustic stimulus on the computer screen. They responded using a binary forced choice experiment interface presented over the Praat demo window function (comparable Praat plugin available at http://www.pholab.uzh.ch/static/volker/software/plugin_BFC_Experiment.zip). After each stimulus presentation, a response window appeared with the question *Französischer oder englischer Akzent?* 'French or English accent?'. Below this text, there were two large grey rectangles titled *Französisch* and *Englisch*. Each of them contained three small blue rectangles that read *sicher* 'confident', *weiss nicht recht* 'not confident', and *nur geraten* 'only guessing'. Listeners clicked on one of the blue rectangles, indicating whether they judged the stimulus as being French- or English-accented German. At the same time, they indicated their confidence level for each stimulus on a 3-point scale. Before the beginning of the experiment, listeners were familiarized with the experiment interface and with manipulated speech through the display of two randomly selected stimuli. The experiment, including instructions, lasted about 20 minutes and listeners were paid 10 Swiss Francs for their participation.

2.6 Data analysis and statistical analyses

Based on listeners' responses, we computed a measure of sensitivity derived from Signal Detection Theory (Green and Swets, 1966) in order to capture listeners' accent identification performance while cancelling out response bias. The non-parametric sensitivity measure A' and the corresponding measure of response bias, B''_D, were calculated following Donaldson (1992). We arbitrarily attributed French-accented German to be signal and English-accented German to be noise; responding 'French' to a French-accented stimulus was thus defined to be a *hit*, whereas responding 'English' to an English-accented stimulus was a *correct rejection*. The two error types, *false alarm* and *miss* were thus the response 'French' to an English-accented stimulus and the response ‘English’ to a French-accented stimulus, respectively. *A’* ranges from 0 to 1, with chance level at 0.5: a listener with an *A’*-value of 0 shows systematic confusion of the stimuli, i.e., responded incorrectly to all stimuli; an *A’*-value of 1 indicates perfect sensitivity. The values for bias (B''_D) range from -1 to 1, 0 indicating no bias, negative values indicating bias towards the response ‘French’ and positive values indicating bias towards ‘English’. An alternative to *A’* and B''_D are the measures *d’* and *β* respectively, which assume underlying normal distributions of hit and false alarm rates. As we obtained comparable results with *d’* and, with one exception (cf. Section 3.2.3), for *β*, we do not report these values. When presenting effects of *accent* and *speaker*, it was not possible to report *A’* as we were interested in the responses to each of the two signal types separately. This is why we reported the percentage of correct responses, *%correct*, instead.

Statistical analyses were performed using R software (R Core Team, 2013). To test the magnitude of listeners’ sensitivity, we calculated two-sided one-sample t-tests. To test for the effect of different factors on listeners’ sensitivity, we constructed linear models (LM). Wherever possible, we calculated linear mixed effects models with *speaker gender*, *accent* and *signal condition* as fixed effects and *speaker*, *sentence* and *listener* as random intercepts (LME; R-package: *lme4*; Bates and Maechler, 2009). We also used linear mixed effect models for acoustic analyses of speech production. Here, our models included *gender*, *accent* and *transplantation* as fixed effects, *speaker* and *sentence* as random intercepts. Effects were tested by comparing a full model, which included the factor in question, to a reduced model, in which the factor was not included. Model comparison was performed using standard likelihood ratio tests (R-code: anova(full_model, reduced_model). We report AIC (Akaike Information Criterion) values for the relative goodness of fit of LMEs (Kliegl et al., 2011). For multiple comparisons, we applied the Tukey method, using the R-package *multcomp*. For correlations, we report Spearman’s correlation coefficient. We assumed an α-level of 0.05.

3 Results

We present results on listeners’ accent identification performance in *timeOnly* and *freqOnly* signal conditions in Section 3.1.1, and Section 3.1.2 compares these results with findings on accent identification performance when both types of cues are combined, in *time+freq* (adapted from Kolly and Dellwo, 2014). In Section 3.2, we investigate potential acoustic correlates of the perceptual results: To verify that our stimuli convey temporal or spectral information of the foreign accents only, we describe the acoustic features of the stimuli in Section 3.2.1. Section 3.2.2 investigates whether acoustic temporal features of the *timeOnly* stimuli may explain listeners’ identification performance. In Section 3.2.3, we explore how the native German segmental content may have biased listeners’ responses in the *timeOnly* condition.

3.1 Results from the perception experiments

3.1.1 Temporal cues and spectral cues in foreign accent identification

To test the magnitude of listeners’ sensitivity, we calculated *A’* for each listener (*n*=40). A boxplot of *A’* for each *signal condition* is presented in Figure 3 (left graph). One-sample t-tests showed that sensitivity was significantly above chance for *timeOnly* (t(19)=2.42, p<0.05*) as well as for *freqOnly* (t(19)=7.69, p<0.001*). We found a significant effect of *condition*: listeners identified accents with greater performance in *freqOnly* (M=0.63, SD=0.08) than in *timeOnly* (M=0.54, SD=0.07; LM: F(1,38)=15.85, p<0.001*).

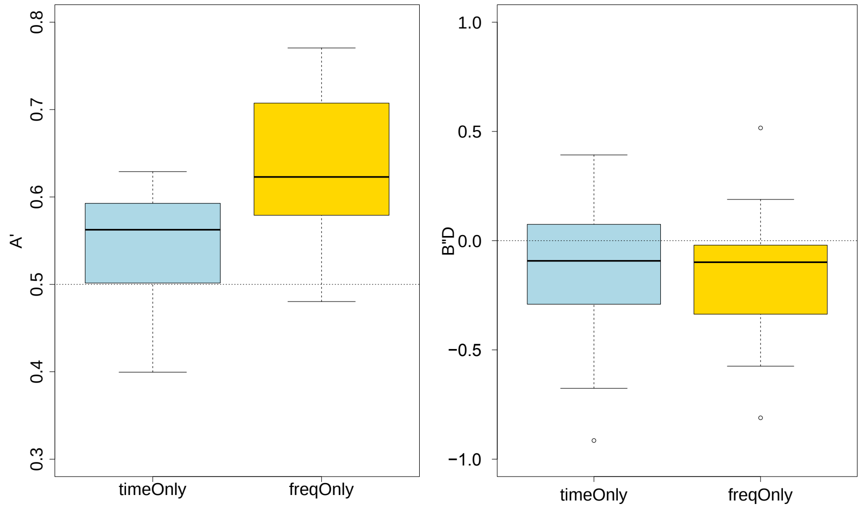

Figure 3: Boxplots of listeners' accent identification performance as measured by *A'* (left graph) and listeners' response bias as measured by B''_D (right graph), by *signal condition*. The dotted lines indicate performance at chance level and no bias, respectively.

Figure 3 (right graph) shows one boxplot of B''_D per *signal condition*, indicating listeners' response bias. One-sample t-tests showed that listeners were significantly biased towards the response 'French' for *freqOnly* (t(19)=-2.40, p<0.05*), but not for *timeOnly* (t(19)=-2.02, p=0.06). Listeners' bias did not differ significantly between *timeOnly* (M=-0.15, SD=0.33) and *freqOnly* (M=-0.16, SD=0.30; LM: F(1,38)=0.01, p=0.93).

To test for the effect of *accent*, we calculated *%correct* for each listener's response to each accent (*n*=80: 2 accents × 40 listeners; as we investigated *accent* effects for each signal condition separately, we performed a Bonferroni-adjustment: 0.05/2=0.025). Boxplots of *%correct* by *accent* and *signal condition* are shown in Figure 4. French accents were identified with significantly higher performance than English accents in *timeOnly* (LM: F(1,38)=7.00, p<0.025*; French: M=0.57, SD=0.10, English: M=0.48, SD=0.11) as well as in *freqOnly* (LM: F(1,38)=7.33, p<0.025*; French: M=0.62, SD=0.10; English: M=0.54, SD=0.10).

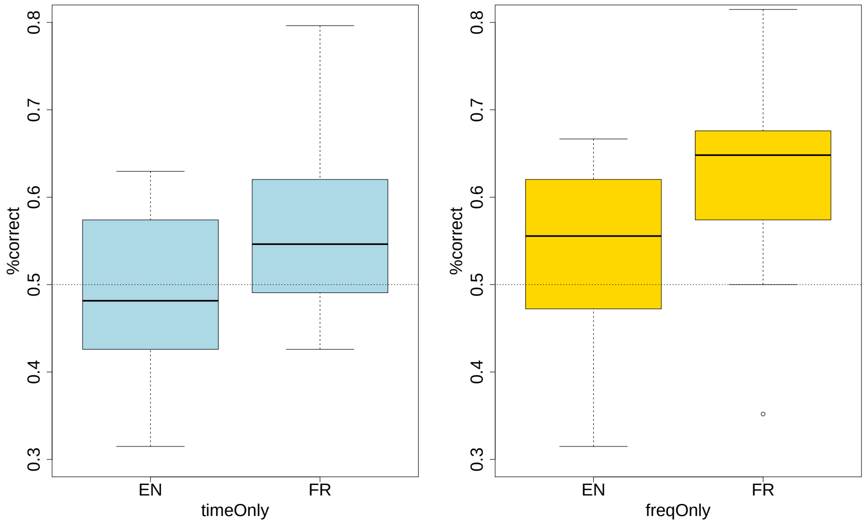

Figure 4: Boxplots of listeners' accent identification performance as measured by *%correct*, per *accent*, for the signal conditions *timeOnly* (left graph) and *freqOnly* (right graph). The dotted lines indicate performance at chance level.

To test for the effect of *speaker*, we calculated *%correct* for each listener's response to each speaker's sentences (*n*=480: 12 speakers × 40 listeners) and constructed an LME of *%correct* with *speaker gender*, *signal condition* and *accent* as fixed effects, a by-*speaker* random slope on *signal condition*, and random intercepts of *speaker* and *listener*. We obtained a significant effect of speaker ($\chi^2(3)=106.91$, AIC=-138.95, p<0.001*). There was no correlation between speakers' strength of foreign accent (cf. Section 2.2.1) and the identification of their accent in either condition (*timeOnly*, r=-0.21, p=0.66; *freqOnly*, r=0.23, p=0.33). To test for the *sentence* effect, we calculated *A'* for each listener's response to each sentence (*n*=720: 18 sentences × 40 listeners) and constructed an LME of *A'* with *signal condition* as fixed effect and random intercepts on *sentence* and *listener*. There was no effect of sentence. Furthermore, listeners' confidence was found not to be significantly affected by *signal condition* (LM: F(1,38)=0.23, p=0.64) or *accent* (LM: F(1,78)=0.83, p=0.37).

3.1.2 Additivity of temporal and spectral cues in foreign accent identification

Figure 5 shows boxplots of *A'* for *timeOnly* (light blue) and *freqOnly* (yellow) in comparison to *time+freq* (green) adapted from Kolly and Dellwo (2014). *Time+freq* contained 6-band noise-vocoded speech with the original, non-native durations, thus featuring both the cues from *timeOnly* and *freqOnly* combined. Results showed a significant overall effect of *condition* (LM: F(2,48)=9.96, p<0.001*). Post-hoc multiple comparisons revealed that *timeOnly* was significantly different from *freqOnly* (p<0.01*) and *time+freq* (p<0.001*); *freqOnly* and *time+freq* did not differ from each other significantly, however (p=0.45). Descriptively, *time+freq* yielded the highest *A'*-values (M=0.67, SD=0.13), followed by *freqOnly* (M=0.63, SD=0.08) and *timeOnly* (M=0.54, SD=0.07).

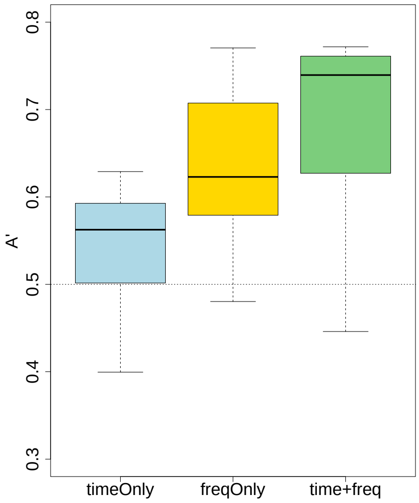

Figure 5: Boxplots of listeners' accent identification performance as measured by *A'*, by *signal condition*, including data from the condition *time+freq* (adapted from Kolly and Dellwo, 2014). The dotted line indicates performance at chance level.

3.2 Results from the acoustic analyses

3.2.1 Acoustic measures of temporal variability in the stimuli

To test whether our material contains the intended acoustical information, we explored which temporal information is contained in the *timeOnly* stimuli, and tested whether *freqOnly* stimuli do in fact contain spectral information alone. To do this, we compared temporal patterns of French- and English-accented German in natural speech and duration transplanted speech. We hereby only applied rhythm metrics of the type (3b), voicing measures, and of the type (3c), peak measures (cf. Section 2.4): when transplanting segment durations, we automatically also copy temporal patterns such as articulation rate, pausing, as well as vocalic and consonantal interval durations. Measures of the type (1)–(3a) are thus not subject to change after duration transplantation. However, when transplanting segment durations to obtain *timeOnly* stimuli, voicing temporal patterns and the location of peaks in the amplitude envelope may only be captured to some extent: The proportion of voicing in individual segments and the location of amplitude peaks are known to differ between languages (voicing: Dellwo et al., 2007; amplitude peaks: Tilsen and Arvaniti, 2013) and speakers (voicing and amplitude peaks: Dellwo et al., 2015; Leemann et al., 2014). We therefore expect these features to be affected by duration transplantation to some extent. Furthermore, non-native speech is often characterized by L1-interference in voicedness, which is why voicing temporal patterns may be a useful cue in the perception task, if French- and English-accented *timeOnly* stimuli were to differ in this feature (Flege and Port, 1981; Hazan and Boulakia, 1993; Leemann, 2011; Neuhauser, 2011; Schmid, 2012; Vieru et al., 2011). In the *freqOnly* stimuli, voicing cues were absent due to 6-band noise-vocoding. However, it is important to examine that the French- and English-accented *freqOnly* stimuli do not differ in amplitude peak durational patterns, as these stimuli are intended to carry spectral cues only.

3.2.1.1 Temporal patterns in timeOnly *stimuli*

Results in Table 2 reveal that four out of five of the applied voicing measures were significantly affected by duration transplantation. Only *varcoVOln* did not differ before and after duration transplantation. The variability of intervals between amplitude peaks, however, seemed to be unaffected by duration transplantation.

Temporal measure	**Factor**	**Result**		
nPVI_VO	*transplantation*	$\chi^2(2)=11.66$	p<0.01*	AIC=1872.1
(voicing measure)	*accent*	$\chi^2(2)=4.42$	p=0.11	AIC=1872.1
	accent ∗ transplantation	$\chi^2(1)=3.84$	p=0.05*	AIC=1872.1
%VO	*transplantation*	$\chi^2(2)=11.38$	p<0.01*	AIC=1423.7
(voicing measure)	*accent*	$\chi^2(2)=5.97$	p=0.05*	AIC=1423.7
	accent ∗ transplantation	$\chi^2(1)=4.25$	p<0.05*	AIC=1423.7
varcoUV	*transplantation*	$\chi^2(2)=10.30$	p<0.01*	AIC=-114.88
(voicing measure)	*accent*	$\chi^2(2)=1.47$	p=0.48	AIC=-114.88
	accent ∗ transplantation	$\chi^2(1)=1.46$	p=0.23	AIC=-114.88
nPVI_UV	*transplantation*	$\chi^2(2)=9.27$	p<0.01*	AIC=1955.5
(voicing measure)	*accent*	$\chi^2(2)=0.87$	p=0.65	AIC=1955.5
	accent ∗ transplantation	$\chi^2(1)=0.74$	p=0.39	AIC=1955.5
varcoPeak	*transplantation*	$\chi^2(2)=1.44$	p=0.49	AIC=-435.61
(peak measure)	*accent*	$\chi^2(2)=0.97$	p=0.62	AIC=-435.61
	accent ∗ transplantation	$\chi^2(1)=0.97$	p=0.33	AIC=-435.61
nPVI_Peak	*transplantation*	$\chi^2(2)=1.08$	p=0.58	AIC=1739.4
(peak measure)	*accent*	$\chi^2(2)=3.39$	p=0.18	AIC=1739.4
	accent ∗ transplantation	$\chi^2(1)=0.95$	p=0.33	AIC=1739.4
varcoVOln	*transplantation*	$\chi^2(2)=0.62$	p=0.73	AIC=-269.11
(voicing measure)	*accent*	$\chi^2(2)=0.23$	p=0.89	AIC=-269.11
	accent ∗ transplantation	$\chi^2(1)=0.08$	p=0.78	AIC=-269.11

Table 2: Summary of the statistics for the tested voicing and peak measures in non-native natural speech and *timeOnly* stimuli. Acoustic measures are ordered according to the magnitude of the effect of *transplantation*.

In the case of *%VO*, we also observed a (marginally) significant effect of *accent* and, for *%VO* as well as *nPVI_VO*, a significant interaction of *transplantation* and *accent*. Simple effects for *%VO* ($\chi^2(1)=10.09$, p<0.01*, AIC=733.7; Bonferroni-adjustment: 0.05/2=0.025) as well as for *nPVI_VO* ($\chi^2(1)=11.61$, p<0.001*, AIC=935.2) showed an effect of *transplantation* in French-accented speech only. Simple effects of *accent* revealed no significant difference between French- and English-accented German in natural or in transplanted speech for neither metric. Figure 6 illustrates a descriptive (but non-significant) difference between voicing temporal patterns of the two accents in natural speech (FR vs. EN, natural), which vanishes in transplanted speech (FR vs. EN, *timeOnly*): *%VO* was higher in French (M=73.95, SD=9.17) than in English (M=69.67, SD=6.91) natural speech; *nPVI_VO* was lower in French (M=66.09, SD=17.67) than in English (M=72.78, SD=17.39) natural speech. Figure 6 further illustrates, for a selection of the durational measures presented in Table 2, that most voicing measures were affected by duration transplantation, whereas the peak measures were not. For example, natural French-accented German exhibits significantly lower values for *nPVI_VO* than duration transplanted French-accented German (natural vs. *timeOnly*, FR). However, there is no such difference regarding the measure *nPVI_Peak*. Based on these results, we conclude that listeners could make little or no use of voicing temporal cues or amplitude peak temporal cues for identifying accents in the signal condition *timeOnly*.

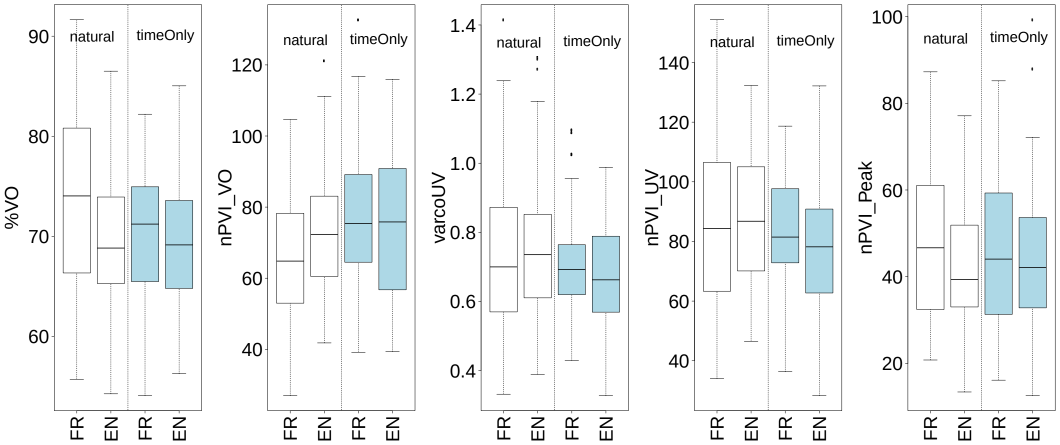

Figure 6: Boxplots of *%VO*, *nPVI_VO*, *varcoUV*, *nPVI_UV* and *nPVI_Peak* in non-native natural speech (white) and in the *timeOnly* stimuli (light blue), for French-accented and English-accented speech.

3.2.1.2 Temporal patterns in freqOnly *stimuli*

As explained in Section 3.2.1, voicing cues are absent from the *freqOnly* stimuli due to noise vocoding. Therefore, only amplitude peak durational measures were applied to these stimuli in order to verify that they contain only frequency domain cues of the foreign accents.

Table 3 shows that neither of the applied peak durational measures in native German speech was significantly affected by duration transplantation. Therefore, *freqOnly* stimuli carry spectral information of the non-native accents, and temporal information of native German, as intended. Furthermore, we found no effect of *accent*, i.e., no difference between French-accented German with native German segment durations and English-accented German with native segment durations. Listeners thus had no durational cues available to complete the perceptual task in the signal condition *freqOnly*.

Temporal measure	**Factor**	**Result**		
varcoPeak (peak measure)	*transplantation*	$\chi^2(2)=3.56$	p=0.17	AIC=-529.36
	accent	$\chi^2(2)=0.19$	p=0.91	AIC=-529.36
	accent ∗ transplantation	$\chi^2(1)=0.17$	p=0.69	AIC=-529.36
nPVI_Peak (peak measure)	*transplantation*	$\chi^2(2)=2.17$	p=0.34	AIC=1681.9
	accent	$\chi^2(2)=0.09$	p=0.96	AIC=1681.9
	accent ∗ transplantation	$\chi^2(1)=0.09$	p=0.77	AIC=1681.9

Table 3: Summary of the statistics for the tested peak measures in native natural speech and *freqOnly* stimuli. Acoustic measures are ordered according to the magnitude of the effect of *transplantation*.

3.2.2 The influence of acoustic measures of temporal variability in foreign accent identification

We calculated correlations of listeners' accent identification performance – as measured by *%correct* – and 16 acoustic measures of temporal variability: two measures of articulation rate (measures of type (1), cf. Section 2.4), two pausing measures (type (2)), five measures of the durational variability of vocalic and consonantal intervals (3a), five measures of the durational

variability of voiced and voiceless intervals (3b) and two measures of the durational variability of intervals between peaks in the amplitude envelope (3c).
Results revealed low correlation coefficients, with |r| ≤ -0.15 for all calculated correlations. Correlation tests were not significant.

3.2.3 The influence of segmental cues in foreign accent identification

We divided the *timeOnly* data into one subset that contained responses to the stimuli featuring uvular /ʀ/s and one subset where uvular /ʀ/s were absent (Bonferroni-adjustment: 0.05/2=0.025). The latter subset contained either no /ʀ/ or vocalized /ʀ/s. Figure 7 shows boxplots of B''_D and *A’* as a function of the presence or absence of uvular /ʀ/s in the stimuli. One-sample t-tests showed that listeners were biased towards the response ‘French’ when the stimuli featured uvular /ʀ/s (t(19)=-5.82, p<0.001*; M=-0.40, SD=0.31), and that they were inclined to answer ‘English’ when no uvular /ʀ/ was present in the stimuli (t(19)=6.31, p<0.001*; M=0.52, SD=0.37). The two subsets significantly differed in bias, as measured by B''_D (LM: F(1,38)=73.50, p<0.001*). However, listeners’ accent recognition performance, as measured by *A’*, did not differ between the two subsets (LM: F(1,38)=0.68, p=0.41).

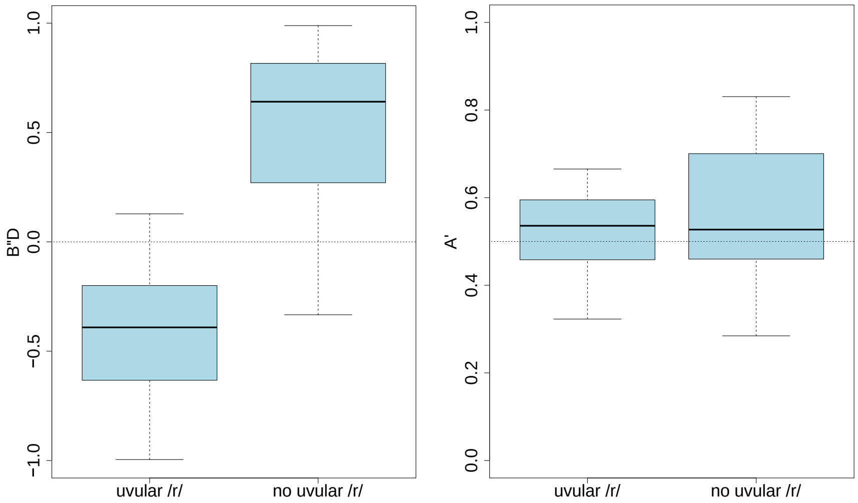

Figure 7: Boxplots of listeners’ accent identification performance as measured by B''_D (left graph) and listeners’ response bias as measured by *A’* (right graph) for stimuli of *timeOnly* which contain uvular /ʀ/s (left boxplot) and which do not contain uvular /ʀ/s (right boxplot).
The dotted lines indicate performance at chance level and no bias, respectively.

4 Discussion

In the present paper, we reported evidence (a) that listeners can, to some extent, identify French- and English-accented German based on temporal features or on degraded spectral features alone, (b) that the combined presence of temporal and spectral cues yields an additive trend towards higher accent identification rates, and (c) that listeners’ response behaviour was biased depending on whether or not stimuli featured uvular /ʀ/s. In the following, we discuss the results obtained in more detail and elaborate on potential implications for forensic phonetics and second language acquisition.

4.1 The importance of temporal and spectral cues in foreign accent identification

4.1.1 Temporal cues

We found that Swiss German listeners could identify speakers’ origin in French- and English-accented German based on temporal information alone in the signal condition *timeOnly*, where segment durations of foreign-accented speech were transplanted to native German speech. In previous experiments, listeners were shown to respond at chance when presented with foreign-accented stimuli that featured temporal information alone, e.g. in monotonized *sasasa*-speech

(Kolly and Dellwo, 2014). However, *sasasa*-speech does not occur in natural situations, whereas the natural speech with manipulated durations used in the present experiment – albeit containing some artifacts – is assumed to sound rather familiar to listeners. This may have enhanced their identification performance, as listeners are used to interpreting temporal information in natural speech, from their everyday life. This is not the case for *sasasa*-speech.
Kolly et al. (2014) had monotonized and lowpass-filtered (<300 Hz) the sentences used for the present experiment and obtained higher accent identification rates than the ones obtained here. On the one hand, lowpass-filtered speech below 300 Hz may still have contained certain segmental cues that boosted listeners' identification performance. On the other hand, lowpass-filtered stimuli contained voicing temporal cues. Our results on the temporal patterns contained in the stimuli revealed a descriptive (but non-significant) difference between natural French- and English-accented German in voicing temporal patterns. However, voicing temporal patterns of both accents became more similar when duration transplantation was applied to create our stimuli (cf. Section 3.2.1). Next to segmental cues below 300 Hz, the additional voicing temporal cues contained in lowpass-filtered speech may therefore account for the different identification performance between the listeners tested by Kolly et al. (2014) and those tested in the present experiment, as it was previously demonstrated that the voicedness of consonants is an important cue in foreign accent identification (Flege and Port, 1981; Vieru et al., 2011). Lowpass-filtered speech also contained the original intensity features of the foreign accents at hand; however, at least their timing should have been very similar between lowpass-filtered and *timeOnly* speech, as duration transplantation was shown not to affect our amplitude peak durational measures (cf. Section 3.2.1).

Which acoustic correlates may account for listeners' sensitivity to temporal cues? We found that accent identification performance did not correlate with any of the applied acoustic measures of temporal variability. Some of these acoustic measures were shown to be affected by duration transplantation, which eliminated (descriptive) differences between French- and English-accented German (cf. Section 3.2.1.1). Considering the low overall correlations, we assume that listeners' response behaviour was driven by patterns of temporal variability not revealed by the temporal measures applied in this study. For example, it may be interesting to investigate patterns of utterance-final lengthening in the future. These have been shown to differ between native and non-native accents of English (White et al., 2012), and to predict adults' and infant's discrimination of accents (White et al, 2012, 2014).

Compared to *A'*-values of 1 for perfect sensitivity, the *A'*-values reported here (M=0.54, SD=0.07) are fairly low. This may be due, to some extent, to the bias driven by the /ʀ/-variants present in the stimuli (cf. Section 3.2.3) and to some of the artifacts contained in our stimuli, which resulted from stretching certain segments and which are likely to be irritating for listeners. However, the sensitivity values reported here are in line with other experiments that use manipulated speech: Ramus et al. (2003), for example, reported mean *A'*-values between 0.57 and 0.74 for listeners' discrimination of languages based on speech temporal cues (undoubtedly, other cues come into play in accent and language identification).

4.1.2 Spectral cues

Degraded spectral features of 6-band noise-vocoded speech were shown to carry enough information for listeners to identify French- and English-accented German above chance, when temporal cues were absent due to duration transplantation for the signal condition *freqOnly* (the absence of temporal cues was demonstrated in Section 3.2.1.2). This is in line with findings by Munro et al. (2010), where listeners could identify native vs. non-native speech in utterances that were played backwards, which also largely disrupts temporal information. These findings emphasize the power of spectral information: Even when speech is strongly degraded in the frequency domain, listeners can process the remaining information, for instance the quality of certain segments, in order to identify foreign accents.

4.1.3 Comparison between results based on temporal and on spectral cues

Listeners' sensitivity to reduced spectral cues in the signal condition *freqOnly* was higher than listeners' sensitivity to segmental temporal cues in *timeOnly*; this, again, emphasizes the prevalence of spectral cues for accent identification tasks.

For both signal conditions, French-accented German was identified with higher performance than English-accented German. This finding is in line with findings by Kolly and Dellwo (2014) and Kolly et al. (2014) for different types of signal-degraded speech containing primarily temporal cues. On the one hand, this may be explained to some extent by the observed tendency for listeners to be biased towards the response 'French'; bias could not be eliminated when calculating the identification performance for each accent separately (*%correct* instead of *A'*). However, there was no significant bias towards 'French' in the *timeOnly* condition. We conclude that temporal patterns of French-accented German may have sounded more salient to our listeners than those of English-accented German. This corroborates suggestions brought forth by studies in the speech rhythm domain: English and German seem to be perceptually more similar in their rhythmic organization, and they differ from French in this regard (Abercrombie, 1967; Dellwo et al., 2007; Grabe and Low, 2002; Pike, 1945; Ramus et al., 1999). Furthermore, this suggests that features of such language-specific temporal patterns are carried over to non-native speech (Arslan and Hansen, 1997; McAllister et al., 2002). Support for this idea was also reported in research by Ordin and Polyanskaya (2015), who found German learners of English to be more successful in acquiring target-like patterns of durational variability than French learners of English.

We found an overall effect of speaker, where some speakers' linguistic origin was identified with higher performance than others'. Non-native speakers thus seem to use different timing strategies when speaking a second language. Temporal features are also known to differ between speakers in their native language (Arvaniti, 2012; Dellwo et al., 2015; Leemann et al., 2014; Wiget et al, 2010). Possibly, speakers' non-native speech may be characterized by similar speaker-idiosyncratic temporal patterns as their native speech, as shown by Kolly et al. (2015) for durational features of silent pauses. Furthermore and interestingly, accent identification scores for each speaker did not correlate with speakers' strength of foreign accent for either signal condition. This may suggest that the information retained in our *timeOnly* and *freqOnly* stimuli was not particularly salient in terms of strength of foreign accent, when listeners judged natural speech. Other features of foreign-accented speech seem to be more important for listeners' perception of accent strength.

4.2 The additivity of temporal and rudimentary spectral cues in foreign accent identification

The combined presence of cues from *timeOnly* as well as *freqOnly* signals in the *time+freq* condition, which contained temporal as well as degraded spectral cues in 6-band noise-vocoded speech, showed a trend towards higher accent identification performance than each type of cue separately. A significant difference was observed between performance in the *timeOnly* vs. *time+freq* condition. The finding is intuitively sound: when the information available to listeners increases, identification performance increases. This is evidence for an additive effect of temporal and spectral cues; however, the combined effect of temporal and spectral cues was smaller than the sum of single effects (Du et al., 2011; Hjalmarsson, 2011). In a similar way, Cunningham-Andersson and Engstrand (1989) have shown that perceived strength of foreign accent increases with the number of target-deviant features. We conclude that the combination of temporal and spectral cues is helpful for listeners to identify foreign accents, but it is not necessary – as each type of cue allowed accent identification above chance on its own. This is also in line with findings by Cunningham-Andersson and Engstrand (1989): some target-deviant features are more strongly associated with the perception of foreign accent than others, and different combinations of such features may increase the perception of accent strength to different degrees.

4.3 The influence of segmental cues in foreign accent identification

We found a significant bias depending on whether or not stimuli featured uvular /ʀ/s. Listeners were biased towards the response 'French' when *timeOnly* stimuli featured uvular /ʀ/s and towards 'English' when they did not. In the *timeOnly* experiment, listeners heard native German segments with French- or English-accented segment durations. However, they were not aware that the segmental content of stimuli was native German; they were only told that they would hear French- and English-accented German. All our Zurich German speakers used uvular /R/ sounds ([ʀ] or [ʁ]), and vocalized /R/s ([ɐ]): the same /R/ sounds as the ones used in Standard German from Germany. But – for the uvular /R/s – these are also the /R/ sounds used in French.

It thus seems that the listeners took the articulation of /R/ as a cue, in a task that was designed to be completed based on durational characteristics alone. Therefore, this affected their response behaviour – and bias – without affecting their accent identification performance. This finding suggests that the duration transplantation method has some pitfalls when used in an identification task design, which is probably less the case when used in an experiment designed to elicit responses on accent strength (Quené and van Delft, 2010; Tajima et al., 1997; Winters and O'Brien, 2013). The finding further stresses the importance of segmental information in foreign accent identification tasks (Boula de Mareüil et al., 2008; Cunningham-Andersson and Engstrand, 1989; Vieru et al., 2011). The articulation of /R/, in particular, seems to be a crucial cue for accent identification in different target languages: Vieru et al. (2011) report it to be one of the most important cues for perceptual foreign accent identification as well as for automatic accent classification. Cunningham-Andersson and Engstrand (1989) found that target-deviant features related to the articulation of /R/ were among the ones that listeners perceived as most accented, whereas target-deviant durational characteristics were amongst the least noticeable. Flege (1984) also cites /R/ as being a strong cue for the detection of (non-)nativeness.

4.4 Possible implications of this work

On the one hand, implications of this research may be found in the domain of forensic phonetics (cf. Section 1): First, the identification of a foreign accent helps narrowing down a group of suspects in forensic casework (speaker profiling; Ellis, 1994; Köster et al., 2012). Since incriminating recordings are most often made over a telephone – the quality of which cannot be controlled for –, temporal features are highly relevant. Second, foreign accent identification is relevant to some LADO cases (Cambier-Langeveld, 2010: 73; Language and National Origin Group, 2004; Verrips, 2011: 137). In LADO, telephone speech is also frequently used (Baltisberger and Hubbuch, 2010). Telephone conditions are one of the reasons for investigating listeners' accent identification performance in speech that contains temporal cues only or reduced spectral cues in general – and therefore for investigating additive effects of temporal and spectral cues in perceptual foreign accent identification.

On the other hand, this research may have implications for the domain of second language acquisition. Speakers who are discriminated against because of their particular accent and origin (Lippi-Green, 1997: 229; Schairer, 1992), for example, might wish to reduce their foreign accent to sound more native-like. It may therefore be helpful to know which accent-specific features are perceptually salient to native listeners. The present experiments suggest that French and English learners of German could take heed of temporal patterns, complementing their regular pronunciation training. Van Santen and Shih (2000) showed that durations of suprasegmental units such as the syllable strongly depend on intrinsic durations of the segments they contain. Therefore, production training focusing on the target-like pronunciation of individual segments, including their durations, may not only improve non-native speakers' production of segmental temporal patterns (e.g. vowel quantity, which is a distinctive feature of German), it could also influence the overall suprasegmental temporal features of their non-native speech towards more native-like productions (Quené and van Delft, 2010; Tajima et al., 1997). Furthermore, the pronunciation of /R/ seems to be a feature worth focusing on if a foreign accent is to be reduced.

5 Summary and conclusion

Our findings showed that listeners could, to a certain extent, identify the linguistic origin of French and English speakers in foreign-accented German, based solely on temporal features of these accents. Furthermore, listeners could also identify the accents in question in stimuli that contain strongly degraded spectral features alone. The combined presence of temporal and spectral information is thus not necessary for listeners to identify foreign accents better than chance. However, we found an additive trend when temporal and spectral cues were combined.

We further found that the segmental information available to listeners biased their response behaviour. When stimuli featured uvular /r/s, listeners were biased towards perceiving a French accent, and a bias towards an English accent was observed in stimuli that featured vocalized or no /r/s. Segmental information – or spectral information – is highly salient and may supress listeners' attention to temporal cues to some extent. Furthermore, the /r/ pronunciation seems to be a very strong cue for listeners to make decisions about a speaker's linguistic origin. However, we found a wide range of acoustic temporal measures not to correlate with listeners' response behaviour. In future work, other measures of temporal variability will have to be explored in order to explain the perceptual results presented here.

The findings may be relevant for forensic phonetics, where particular cues of foreign-accented speech allow practitioners or ear-witnesses to identify a speaker's linguistic origin – and where advice often has to be given based on speech that is degraded by telephone networks or background noise. Our findings may also have implications for second language acquisition. Some non-native speakers may wish to reduce their foreign accent. In such cases, it is crucial to know which features of an accent are perceptually salient to native listeners.

Acknowledgements

This research was supported by the Swiss National Science Foundation (SNSF; grant numbers P1ZHP1_155024 and 100015_135287). The authors would like to thank their subjects, speakers as well as listeners, for their contribution to these experiments. Furthermore, they thank Stephan Schmid for his expert advice on foreign-accented speech, Camilla Bernardasci for her help with some of the recordings and Andrea Fröhlich for carrying out the perception experiment with some of the listeners. The present paper benefited from the helpful comments of two anonymous reviewers and the section editor, whom we would also like to thank.

Appendix: Reading materials

01 Die Frau des Apothekers weiss immer, was sie will.
02 Das Theater hat viele neue Aufführungen geplant.
03 Er wollte sich seiner Schwächen einfach nicht bewusst werden.
04 Der öffentliche Verkehr lässt viel zu wünschen übrig.
05 Die schlechte Zahlungsbilanz lässt mich nicht zur Ruhe kommen.
06 Die Eltern geben ihm keine finanzielle Unterstützung.
07 Der starke Frühlingsregen hat grossen Schaden angerichtet.
08 Der schnellste Zug ist immer noch der ICE.
09 Der Wiederaufbau der Stadt wird sehr lange dauern.
10 Das Bildungsministerium hat den einfachsten Weg gewählt.
11 Diese Konditorei macht ausgezeichnete Kuchen.
12 Dieses Geschäft bietet sehr preisgünstige Ware an.
13 Sie haben die Wahrheit erst entdeckt, als er auspackte.
14 Für meine Mannschaft wird der Sieg ein Kinderspiel sein.
15 Die Meinungsumfragen sagen einen Sieg der Rechten voraus.
16 Die Strassen der Innenstadt wurden von der Polizei gesperrt.
17 Ein berühmtes Bild wurde aus dem Kunsthaus gestohlen.
18 Der Müssiggang ist bekanntlich aller Laster Anfang.

References

Abercrombie, D., 1967. Elements of general phonetics. Edinburgh, Edinburgh University Press.

Arslan, L. M., Hansen, J. H., 1997. A study of temporal features and frequency characteristics in American English foreign accent. Journal of the Acoustical Society of America 102 (1), 28–40.

Arvaniti, A., 2012. The usefulness of metrics in the quantification of speech rhythm. Journal of Phonetics 40, 351–371.

Auer, P., 2001. Silben- und akzentzählende Sprachen, in: Haspelmath, M., König, E., Oesterreicher, W. (Eds), Language typology and language universals. An international handbook. Vol. 2. Berlin/New York, de Gruyter, pp. 1391–1399.

Baltisberger, E., Hubbuch, P., 2010. LADO with specialized linguists – The development of LINGUA's working method, in: Zwaan, K., Verrips, M., Muysken, P. (Eds.), Language and origin: The role of language in European asylum procedures. Nijmegen, Wolf Legal Publishers, pp. 9–19.

Bates, D. M., Maechler, M., 2009. lme4: Linear mixed-effects models using S4 classes. R package version 1.1-7.

Boersma, P., Weenink, D., 2014. Praat: Doing phonetics by computer. Version 5.4. http://www.praat.org/.

Bosker, H. R., Quené, H., Sanders, T., Jong, N. H., 2014. The perception of fluency in native and nonnative speech. Language Learning 64, 579–614.

Boula de Mareüil, P., Brahimi, B., Gendrot, C., 2004a. Role of segmental and suprasegmental cues in the perception of Maghrebian-accented French. Proceedings of the International Conference on Spoken Language Processing 2004, Jeju, pp. 341–344.

Boula de Mareüil, P., Marotta, G., Adda-Decker, M., 2004b. Contribution of prosody to the perception of Spanish/Italian accents. Proceedings of Speech Prosody 2004, Nara.

Boula de Mareüil, P., Vieru-Dimulescu, B., 2006. The contribution of prosody to the perception of foreign accent. Phonetica 63 (4), 247–267.

Boula de Mareüil, P., Vieru-Dimulescu, B., Woehrling, C., Adda-Decker, M., 2008. Accents étrangers et régionaux en français. Traitement Automatique des Langues 49 (3), 135–163.

Byrne, C., Foulkes, P., 2004. The 'mobile phone effect' on vowel formants. Journal of Speech, Language and the Law 11 (1), 83–102.

Cambier-Langeveld, T., 2010. The role of linguists and native speakers in language analysis for the determination of speaker origin. Journal of Speech, Language and the Law 17 (1), 67–93.

Chen, B., Zhu, Q., Morgan, N., 2005. Long-term temporal features for conversational speech recognition, in: Bengio, S., Bourlard, H. (Eds.), Machine learning for multimodal interaction. Berlin/Heidelberg/New York, Springer, pp. 232–242.

Council of Europe, 2013. Common European framework of reference for languages: Learning, teaching, assessment. http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf (accessed 12.10.2015).

Cucchiarini, C., Strik, H., Boves, L., 2002. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. Journal of the Acoustical Society of America 111, 2862–2873.

Cunningham-Andersson, U., Engstrand, O., 1989. Perceived strength and identity of foreign accent in Swedish. Phonetica 46, 138–154.

Dauer, R. M., 1983. Stress-timing and syllable-timing reanalyzed. Journal of Phonetics 11, 51–62.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., McGettigan, C., 2005. Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. Journal of Experimental Psychology 134 (2), 222–241.

Delattre, P., 1966. A comparison of syllable length conditioning among languages. International Review of Applied Linguistics in Language Teaching 4 (3), 183–198.

Dellwo, V., 2006. Rhythm and speech rate: A variation coefficient for DeltaC, in: Karnowski, P., Szigeti, I. (Eds.), Language and language-processing. Frankfurt am Main, Lang, pp. 231–241.

Dellwo, V., 2008. The role of speech rate in perceiving speech rhythm. Proceedings of Speech Prosody 2008, Campinas, pp. 375–378.

Dellwo, V., 2010. Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence. PhD Thesis, University of Bonn.

Dellwo, V., Fourcin, A., Abberton, E., 2007. Rhythmical classification of languages based on voice parameters. Proceedings of the International Congress of Phonetic Sciences 2007, Saarbrücken, pp. 1129–1132.

Dellwo, V., Leemann, A., Kolly, M.-J., 2012. Speaker idiosyncratic rhythmic features in the speech signal. Proceedings of Interspeech 2012, Portland, pp. 1584–1587.

Dellwo, V., Leemann, A., Kolly, M.-J., 2015. Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. Journal of the Acoustical Society of America 137 (3), 1513–1528.

Donaldson, W., 1992. Measuring recognition memory. Journal of Experimental Psychology: General 121 (3), 275–277.

Du, Y., He, Y., Ross, B., Bardouille, T., Wu, X., Li, L., Alain, C., 2011. Human auditory cortex activity shows additive effects of spectral and spatial cues during speech segregation. Cerebral Cortex 21 (3), 698–707.

Ellis, S., 1994. The Yorkshire Ripper enquiry: Part 1. Forensic Linguistics 1, 197–206.

Erziehungsdirektion des Kantons Bern, 2009. Sprachniveau an der Maturität gemäss Europäischem Sprachenportfolio (ESP). http://www.erz.be.ch/erz/de/index/mittelschule/mittelschule/publikationen.assetref/dam/documents/ERZ/MBA/de/AMS/ams_sprachniveau_maturitaet.pdf, accessed 05.05.2016).

Fant, G., Kruckenberg, A., Nord, L., 1991. Durational correlates of stress in Swedish, French and English. Journal of Phonetics 19 (3–4), 351–365.

Ferguson, C. A., 1959. Diglossia. Word 15, 325–340.

Ferragne, E., Pellegrino, F., 2004. Rhythm in read British English: Interdialect variability. Proceedings of the International Conference on Spoken Language Processing 2004, Jeju, pp. 1573–1576.

Flege, J. E., 1984. The detection of French accent by American listeners. Journal of the Acoustical Society of America 76 (3), 692–707.

Flege, J. E., Port, R., 1981. Cross-language phonetic interference: Arabic to English. Language and Speech 24 (2), 125–146.

Grabe, E., Low, E. L., 2002. Durational variability in speech and the Rhythm Class Hypothesis, in: Gussenhoven, C., Warner, N. (Eds.), Laboratory Phonology. Berlin/New York, Mouton de Gruyter, pp. 515–545.

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

Grenon, I., White, L., 2008. Acquiring rhythm. A comparison of L1 and L2 speakers of Canadian English and Japanese. Proceedings of the Boston University Conference on Language Development 2008, Boston, pp. 155–166.

Hazan, V. L., Boulakia, G., 1993. Perception and production of a voicing contrast by French-English bilinguals. Language and Speech 36 (1), 17–38.

Hirson, A., French, P., Howard, D., 1995. Speech fundamental frequency over the telephone and face-to-face: Some implications for forensic phonetics, in: Windsor Lewis, J. (Ed.), Studies in general and English phonetics in honour of Professor J.D. O'Connor. London, Routledge, pp. 230–240.

Hjalmarsson, A., 2011. The additive effect of turn-taking cues in human and synthetic voice. Speech Communication 53 (1), 23–35.

Holm, S., 2008. Intonational and durational contributions to the perception of foreign-accented Norwegian. An experimental phonetic investigation. PhD Thesis, Norwegian University of Science and Technology.

Hove, I., 2002. Die Aussprache der Standardsprache in der deutschen Schweiz. Berlin/New York, de Gruyter.

de Jong, N. H., Groenhout, R., Schoonen, R., Hulstijn, J. H., 2013. Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behaviour. Applied Psycholinguistics 34, 1–21.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., Zhou, X., 2011. Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. Frontiers in Psychology 1, 1–12.

Kohler, K., 1990. German. Journal of the International Phonetic Association 20, 48–50.

Kolde, G., 1981. Sprachkontakte in gemischtsprachigen Städten. Vergleichende Untersuchungen über Voraussetzungen und Formen sprachlicher Interaktion verschiedensprachiger Jugendlicher in den Schweizer Städten Biel/Bienne und Fribourg/Freiburg i. Ue. Wiesbaden, Steiner.

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

Kolly, M.-J., Leemann, A., Dellwo, V., 2014. Foreign accent recognition based on temporal information contained in lowpass-filtered speech. Proceedings of Interspeech 2014, Singapore, pp. 2175–2179.

Kolly, M.-J., Leemann, A., Boula de Mareüil, P., Dellwo, V., 2015. Speaker-idiosyncrasy in pausing behavior: Evidence from a cross-linguistic study. Proceedings of the International Congress of Phonetic Sciences 2015, Glasgow.

Köster, O., Kehrein, R., Masthoff, K., Boubaker, Y. H., 2012. The tell-tale accent: Identification of regionally marked speech in German telephone conversations by forensic phoneticians. Journal of Speech, Language and the Law 19 (1), 51–71.

Künzel, H. J., 2001. Beware of the 'telephone effect'. The influence of telephone transmission on the measurement of formant frequencies. Forensic Linguistics 8 (1), 80–99.

Künzel, H. J., 1997. Some general phonetic and forensic aspects of speaking tempo. Journal of Speech Language and the Law 4, 48–83.

Laeufer, C., 1992. Patterns of voicing-conditioned vowel duration in French and English. Journal of Phonetics 20 (4), 411–440.

Language and National Origin Group, 2004. Guidelines for the use of language analysis for the determination of the origin of asylum seekers. Journal of Speech, Language and the Law 16 (1), 113–138.

Leemann, A., 2011. Einfluss der Schweizerdeutschen Phonologie auf die Stimmhaftigkeit von Frikativen im L2-Englischen. Poster presented at the 'Phonetik und Phonologie' Conference 2011, Osnabrück.

Leemann, A., Dellwo, V., Kolly, M.-J., Schmid, S., 2012. Rhythmic variability in Swiss German dialects. Proceedings of Speech Prosody 2012, Shanghai, 607–610.

Leemann, A., Kolly, M.-J., Dellwo, V., 2014. Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. Forensic Science International 238, 59–67.

Lippi-Green, R., 1997. English with an accent: Language ideology and discrimination in the United States. London/New York, Routledge.

Lloyd James, A., 1929. Historical introduction to French phonetics. London, University of London Press.

Maassen, B., Povel, D.-J., 1985. The effect of segmental and suprasegmental corrections on the intelligibility of deaf speech. Journal of the Acoustical Society of America 78 (3), 877–886.

McAllister, R., Flege, J. E., Piske, T., 2002. The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. Journal of Phonetics 30 (2), 229–258.

Mermelstein, P., 1975. Automatic segmentation of speech into syllabic units. Journal of the Acoustical Society of America 58 (4), 880–883.

Munro, M. J., Derwing, T. M., Burgess, C. S., 2010. Detection of nonnative speaker status from content-masked speech. Speech Communication 52 (7), 626–637.

Nazzi, T., Bertoncini, J., Mehler, J., 1998. Language discrimination by newborns: Toward an understanding of the role of rhythm. Journal of Experimental Psychology: Human Perception and Performance 24 (3), 756–766.

Neuhauser, S., 2011. Foreign accent imitation and variation of VOT and voicing in plosives. Proceedings of the International Congress of Phonetic Sciences 2003, Barcelona, pp. 1462–1465.

Ordin, M., Polyanskaya, L., 2015. Acquisition of speech rhythm in a second language by learners with rhythmically different native languages. Journal of the Acoustical Society of America, 138, 533–544.

Osberger, M. J., Levitt, H., 1979. The effect of timing errors on the intelligibility of deaf children’s speech. Journal of the Acoustical Society of America 66 (5), 1316–1324.

Pettorino, M., Vitale, M., 2012. Transplanting native prosody into second language speech, in: Busà, M. G., Stella, A. (Eds.), Methodological perspectives on second language prosody. Padova, Coop. Libraria Editrice Università di Padova, pp. 11–16.

Pike, K., 1945. The intonation of American English. Ann Arbor, University of Michigan Press.

Pinet, M., Iverson, P., 2010. Talker-listener accent interactions in speech-in-noise recognition: Effects of prosodic manipulation as a function of language experience. Journal of the Acoustical Society of America 128 (3), 1357–1365.

Quené, H., van Delft, L. E., 2010. Non-native durational patterns decrease speech intelligibility. Speech Communication 52, 911–918.

Ramus, F., Mehler, J., 1999. Language identification with suprasegmental cues: A study based on speech resynthesis. Journal of the Acoustical Society of America 105 (1), 512–521.

Ramus, F., Nespor, M., Mehler, J., 1999. Correlates of linguistic rhythm in the speech signal. Cognition 73, 265–292.

Ramus, F., Dupoux, E., Mehler, J., 2003. The psychological reality of rhythm classes. Perceptual studies. Proceedings of the International Congress of Phonetic Sciences 2003, Barcelona, pp. 337–342.

R Core Team, 2013. R. A language and environment for statistical computing. Version 3.0.1. Vienna. http://www.R-project.org.

Rognoni, L., Busà, M. G., 2014. Testing the effects of segmental and suprasegmental phonetic cues in foreign accent rating: An experiment using prosody transplantation. Proceedings of the International Symposium on the Acquisition of Second Language Speech 2013, Montreal, pp. 547–560.

Schairer, K. E., 1992. Native speaker reaction to non-native speech. Modern Language Journal 76 (3), 309–319.

Schmid, S., 2012. The pronunciation of voiced obstruents in L2 French: A preliminary study of Swiss German learners. Poznań Studies in Contemporary Linguistics 48, 627–659.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–304.

Shearme, J. N., Holmes, J. N., 1961. An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1 - formant 2 plane. Proceedings of the International Congress of Phonetic Sciences 1961, Helsinki.

Tajima, K., Port, R., Dalby, J., 1997. Effects of temporal correction on intelligibility of foreign-accented English. Journal of Phonetics 25 (1), 1–24.

Tiffany, W. R., 1959. Nonrandom sources of variation in vowel quality. Journal of Speech and Hearing Research 2, 305–317.

Tilsen, S., Arvaniti, A., 2013. Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. Journal of the Acoustical Society of America 134 (1), 628–639.

Trofimovich, P., Baker, W., 2006. Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. Studies in Second Language Acquisition 28, 1–30.

van Santen, J. P. H., Shih, C., 2000. Suprasegmental and segmental timing models in Mandarin Chinese and American English. Journal of the Acoustical Society of America 107 (2), 1012–1026.

Van Zyl, M., Hanekom, J. J., 2011. Speech perception in noise: A comparison between sentence and prosody recognition. Journal of Hearing Science 1 (2), 54–56.

Verrips, M., 2011. LADO and the pressure to draw strong conclusions. Journal of Speech Language and the Law 18 (1), 131–143.

Vieru, B., Boula de Mareüil, P., Adda-Decker, M., 2011. Characterisation and identification of non-native French accents. Speech Communication 53 (3), 292–310.

Vitale, M., Boula de Mareüil, P., De Meo, M., 2014. An acoustic-perceptual approach to the prosody of Chinese and native speakers of Italian based yes/no questions. Proceedings of Speech Prosody 2014, Shanghai, pp. 648–652.

Werlen, I., 1980. R im Schweizerdeutschen. Zeitschrift für Dialektologie und Linguistik 47, 52–76.

White, L., Mattys, S. L., 2007a. Calibrating rhythm: First language and second language studies. Journal of Phonetics 35, 501–522.

White, L., Mattys, S. L., 2007b. Rhythmic typology and variation in first and second languages, in: Prieto, P., Mascaró, J., Solé, M.-J. (Eds.), Segmental and prosodic issues in romance phonology. Amsterdam/Philadelphia, Benjamins, pp. 237–257.

White, L., Mattys, S. L., Wiget, L., 2012. Language categorization by adults is based on sensitivity to durational cues, not rhythm class. Journal of Memory and Language 66 (4), 665–679.

White, L., Floccia, C., Goslin, J., Butler, J., 2014. Utterance-final lengthening is predictive of infants’ discrimination of English accents. Language Learning 64 (S2), 27–44.

Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., Mattys, S. L., 2010. How stable are acoustic metrics of contrastive speech rhythm? Journal of the Acoustical Society of America 127 (3), 1559–1569.

Winters, S., O’Brien, M. G., 2013. Perceived accentedness and intelligibility. The relative contributions of f0 and duration. Speech Communication 55, 486–507.