Speech Communication 86 (2017) 121–134

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

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]

[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

ARTICLE INFO

Article history:
Received 13 December 2015
Revised 12 November 2016
Accepted 22 November 2016
Available online 30 November 2016

Keywords:
Accent identification
Speech rhythm
Temporal cues
Segmental cues
Additivity of cues
Forensic phonetics

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 /r/s and towards 'English' when they contained vocalized /r/s or lacked /r/.

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

[*] Corresponding author at: Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, Plattenstrasse 54, 8032 Zurich, Switzerland.

E-mail addresses: marie-jose.kolly@uzh.ch (M.-J. Kolly), philippe.boula.de.mareuil@limsi.fr (P. Boula de Mareüil), al764@cam.ac.uk (A. Leemann), volker.dellwo@uzh.ch (V. Dellwo).

http://dx.doi.org/10.1016/j.specom.2016.11.006
0167-6393/© 2016 Elsevier B.V. All rights reserved.

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; 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;