Title: Detection of Non-native Speaker Status from Backwards and Vocoded Content-masked Speech

Author: Arkadiusz Rojczyk, Andrzej Porzuczek

Citation style: Rojczyk Arkadiusz, Porzuczek Andrzej. (2020). Detection of Non-native Speaker Status from Backwards and Vocoded Content-masked Speech. "Theory and Practice of Second Language Acquisition" (2020), vol. 6 (2), s. 87-105. DOI: 10.31261/TAPSLA.7714

Theory and Practice of Second Language Acquisition
vol. 6 (2), 2020, pp. 87–105
https://doi.org/10.31261/TAPSLA.7714

Arkadiusz Rojczyk https://orcid.org/0000-0002-7328-5911
Acoustic-Phonetic Laboratory, University of Silesia in Katowice
Andrzej Porzuczek https://orcid.org/0000-0001-6398-2150
Acoustic-Phonetic Laboratory, University of Silesia in Katowice

Detection of Non-native Speaker Status from Backwards and Vocoded Content-masked Speech

Abstract

This paper addresses the issue of speech rhythm as a cue to non-native pronunciation. In natural recordings, it is impossible to disentangle rhythm from segmental, subphonemic or suprasegmental features that may influence nativeness ratings. However, two methods of speech manipulation, that is, backwards content-masked speech and vocoded speech, allow the identification of native and non-native speech in which segmental properties are masked and become inaccessible to the listeners. In the current study, we use these two methods to compare the perception of content-masked native English speech and Polish-accented speech. Both native English and Polish-accented recordings were manipulated using backwards masked speech and 4-band white-noise vocoded speech. Fourteen listeners classified the stimuli as produced by native or Polish speakers of English. Polish and English differ in their temporal organization, so, if rhythm is a significant contributor to the status of non-native accentedness, we expected an above-chance rate of recognition of native and non-native English speech. Moreover, backwards content-masked speech was predicted to yield better results than vocoded speech, because it retains some of the indexical properties of speakers. The results show that listeners are unable to detect non-native accent in Polish learners of English from backwards and vocoded speech samples.

Keywords: accent detection, non-native accent, content-masked speech, vocoded speech, backwards speech

Introduction

Non-native speech is usually easily detected not only by native speakers but also by most non-native speakers of a language. Accent identification may

be important for legal (e.g., forensic analyses in speaker identification), sociological and pedagogical reasons. Previous research on accentedness in a non-native language has shown that most EFL learners (e.g., Waniek-Klimczak, Porzuczek, & Rojczyk, 2013) as well as L2 learners prefer to suppress foreign accent traces in order to approach native-like pronunciation patterns. Their motivation may range from signalling higher language competence in their speech to avoiding problems or even discriminatory attitudes in some language communities (Anisfeld, Bogo, & Lambert, 1962; Arthur, Farrar, & Bradford, 1974; Lippi-Green, 1997; Ryan & Carranza, 1975; Schairer, 1992). Suppressing heavy foreign accent also helps speakers to appear more credible to listeners (Lev-Ari & Keysar, 2010), which improves interpersonal communication in both professional and private affairs. Although non-native accent detection is a relatively easy task, it is not clear how various individual cues contribute to its recognition. Such knowledge may help EFL learners to approach the process of learning pronunciation in a more systematic way.

There is evidence that segmental/subphonemic deviations from native speech, such as substitutions, insertions or deletions (see Munro, Derwing, & Burgess, 2010) are the primary cues to foreign accent (Flege & Port, 1981; Kolly, Boula de Mareüil, & Dellwo, 2017). Prosody is also an important factor investigated by researchers, especially intonation, despite its variability across speakers (Mennen, 2004; Trofimovich & Baker, 2006). The temporal properties of speech, or rhythm, form another individually variable cue to accent identification (Tajima, Port, & Dalby, 1997; White & Mattys, 2007). Fluency is indicated by Riggenbach (1991) and Derwing, Munro, and Thomson (2008), while Raupach (1980) points out to articulatory rate, which is supported by Munro and Derwing (2001), who found that digitally accelerated speech is rated as more native-like. Finally, there are also extralinguistic parameters of speech that are said to facilitate accent recognition, such as voice quality (Laver, 1980), related to long-term laryngeal and supralaryngeal setting, which is often carried over from L1 articulatory habits (Esling, 2000; Wilson, 2006) as well as and ethnic vocal tract differences (Andrianopoulos, Darrow, & Chen, 2001).

Needless to say, the listener also has access to the lexical and grammatical structures used by the speaker, which may alone clearly indicate their native or non-native status. For this reason, researchers use content-masked speech, such as vocoded and backwards speech in order to investigate individual cues to foreign accent. Backwards speech retains the temporal properties of the sample in terms of syllable length variation as well as voice quality, intonation (pitch variation), and rhythm understood as the timing relations between prosodic units (Black, 1973; Van Lancker, Kreiman, & Emmorey, 1985; Ramus et al., 2000; Toro, Trobalon, & Sebastián-Gallés, 2003; Munro et al., 2010). With regard to segmental information, according to Black (1973), fricatives and

nasals are perceptible, unlike glides or laterals, or vowel quality in general. Vocoded speech, in turn, displays intensity variation, with the salient peaks corresponding to syllable nuclei (Kolly & Dellwo, 2013; Kolly et al., 2017). The rhythm of utterances is thus observable as the temporal distribution of intensity peaks. Vocoded speech retains no intonation in terms of pitch variation and there remain scarce spectral characteristics of speech, depending on the vocoding parameters. Table 1 compares features provided by backwards, vocoded, and natural speech.

Table 1.

Potential foreign accent cues in two types of content-masked speech

Speech characteristics	Backwards	Vocoded	Natural
Segmental	degraded	no	yes
Pitch variation	yes	no	yes
Intonation	no	no	yes
Vowel duration	yes	no	yes
Phrasal rhythm	no	yes	yes
Voice quality	yes	no	yes

Previous Studies

In this section we report the findings provided by recent studies using the two methods. The discussion provided by Munro et al. (2010) suggests that the speech characteristics observable in backwards-masked speech are usually sufficient for the identification of familiar voices and may also help listeners recognize a foreign accent. The results of their experiments show above chance levels of foreign (Mandarin, Cantonese, and Czech) accent detection in backwards speech samples. These experiments have also shown a moderate effect of speech rate. The influence of pitch has not been ruled out although no statistically significant differences were observed for monotonized and randomly-spliced backwards speech samples, which made it impossible for listeners to use pitch and temporal properties as possible cues to foreign accent. The authors also admit that the results were not sufficient to assess the role of individual voice quality in accent evaluation.

Kolly and Dellwo (2013) and Kolly et al. (2017) investigated how temporal and rhythmic cues alone can be used to identify French- and English-accented German speech in a number of tasks, including *sasasa*-speech, 1-bit requantized speech, 6-band noise vocoded samples and in recordings where native segments

were transplanted into sentences featuring non-native speech unit timing. The authors found that only monotone *sasasa*-speech, displaying the timing of alternate voiced and voiceless speech intervals alone (Fourcin & Dellwo, 2009) was insufficient for listeners to make accent judgments, while they performed above-chance levels in tasks using the 1-bit-quantized and noise vocoded samples. The latter, only offering listeners access to the temporal characteristics of syllable beats, proved to be more problematic than the former, where the listeners were able to perceive segment durations.

The Current Study

This study deals with the detection of Polish-accented content-masked speech. The content is masked by means of (1) reversing the acoustic signal and (2) four-band white-noise vocoding of the recorded samples. In particular, we want to find out whether native English and Polish listeners can detect Polish-accented speech on the basis of temporal cues or rhythm alone. Rhythm is operationalized in our experiment as vowel length variation which can be measured in natural and backwards speech as rate-normalized standard deviation from mean vowel duration using VarcoV (White & Mattys, 2007) and a similar measure, Vowel Reduction Quotient (VRQ), calculated individually by dividing the speakers mean unstressed vowel duration by mean stressed vowel duration (Porzuczek, 2012). Obviously, both VarcoV and VRQ values are identical for natural and backwards speech.

Vocoded speech samples do not reveal vowel duration but it is possible to measure the time intervals between consecutive syllable peaks (vowel mid-points), still perceptible to listeners, and calculate their variability, using VarcoPeak (Dellwo, 2012). Thus manifested temporal speech organization is the only auditory cue that the listeners may rely upon in their judgments.

Backwards speech, though characterized by the same VarcoPeak quotients, cannot be rated with respect to rhythm since the reversed prominence distribution becomes meaningless. In effect, in vocoded speech we can observe the rhythm roughly understood as syllable length variation, but this property is inaccessible in backwards speech, which, in turn, features vowel length variation, pitch variation and range, and some spectral (segmental) information.

Following the results from previous studies, we hypothesize that with respect to predicted ceiling efficiency of Polish-accented speech recognition, listeners may still show an above-chance level of judgment efficiency in the case of both types of content-masked speech samples.

Materials

The speakers were three male native users of Standard Southern British English (SSBE), aged approximately 20, 40, and 60 (E1, E2, E3 respectively), and three Polish second-year students of English (P1, P2, P3) aged 20, one female and two males, recruited at the Institute of English, University of Silesia in Katowice, Poland. Their proficiency, confirmed by regular curriculum tests, was B2 to C1 in the Common European Framework of Reference for Languages (CERFL). The learners' target pronunciation model was also SSBE, typical of Polish EFL education system. They had detectable segmental, rhythmic, and prosodic features of Polish-accented English. All speakers were asked to read a short passage of 124 words describing theft in a shop (Alexander, 1967). The recording took place in a sound-proof booth in the Acoustic-Phonetic Laboratory at the University of Silesia. The signal was captured at 44,100 Hz (24-bit quantization) through a headset dynamic microphone Sennheiser HMD 26 fed by a USBPre2 (Sound Devices) amplifier.

The test materials included three phrases extracted from the passage:

– *that the shop assistant was her daughter* (10 syllables; Mean duration = 1,558 ms);
– *she chose one of the most expensive dresses in the shop* (14 syllables; Mean duration = 2,474 ms);
– *the temptation to steal is greater than ever before* (14 syllables; Mean duration 2,682 ms).

These phrases were selected, because their syntactic and prosodic structure provided contexts for the greatest expected rhythmical variability between native and Polish-accented productions. More specifically, they include several strings of two and three consecutive reduced syllables, which strongly contribute to the prototypical English stress-timing. The phrases were normalized for intensity (70 dB) and duration. The mean phrase duration was calculated for all speakers and each phrase was digitally stretched or compressed to the mean using Pitch Synchronous Overlap and Add (PSOLA) in Praat (Boersma, 2001). Importantly, although PSOLA temporal manipulation alters raw durations of individual segments, it retains their proportional durations, thus maintaining the rhythmical structure of the phrases. At the same time, it rules out the possibility that slower productions will be rated as non-native irrespective of their rhythmical properties, because speaking rate has been shown to significantly influence native/non-native accent ratings (Munro & Derwing, 2001).

The backwards-masked phrases were created by digitally reversing the natural phrases, using the 'reverse selection' in Praat. The vocoded phrases were created using Praat script. Noise vocoding relies on extracting amplitude envelopes from several frequency bands to modulate white noise in those

frequencies (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Most of the spectral information critical for accent identification, such as vowel quality or voicing, is degraded or completely absent and listeners have access only to speech rhythm represented by syllable beats in the form of white noise pulses (Cummings & Port, 1998; Kolly & Dellwo, 2014; Lee & Todd, 2004; Tilsen & Arvaniti, 2013). The actual degree of spectral degradation depends on the number of bands in vocoding. As reported by Kolly and Dellwo (2014), 6-band noise vocoded speech retains some spectral information that allows discrimination of vowels. On the other hand, 3-band noise vocoded speech degrades all spectral information that may facilitate identification of individual segments. Importantly, Kolly et al. (2017) showed that 6-band noise vocoded speech carries enough spectral information for listeners to identify French- and English-accented German above chance even if durational cues are absent due to duration transplantation. On the other hand, 3-band noise vocoding did not allow any identification of French- and English-accented German from durational cues (Kolly & Dellwo 2014). As a result, we decided to use intermediate 4-band noise vocoding of the test phrases. The perceptual impression was a sequence of white noise pulses without any clearly identifiable properties of vowels and consonants. Figures 1 and 2 show the natural and vocoded phrase *that the shop assistant was her daughter.*

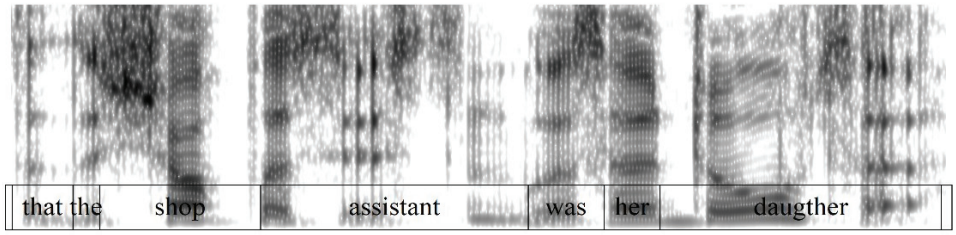

Figure 1. Spectrogram of the natural phrase *that the shop assistant was her daughter.*

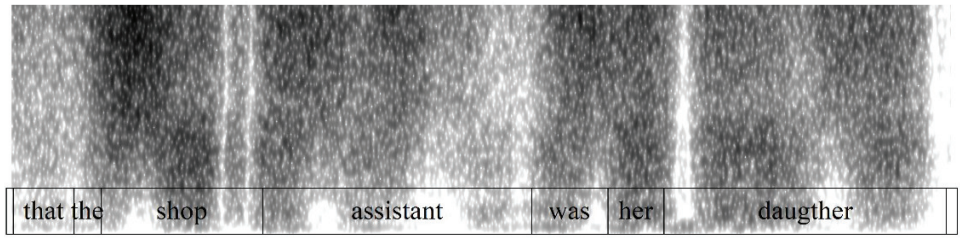

Figure 2. Spectrogram of the 4-band noise vocoded phrase *that the shop assistant was her daughter.*

Using VarcoV and VRQ for natural and backwards speech, we indicated the different rhythmic tendencies in native and Polish-accented speech. These are illustrated in Figures 3–5 below. VRQ figures have been multiplied by 100 for more direct comparison with VarcoV.

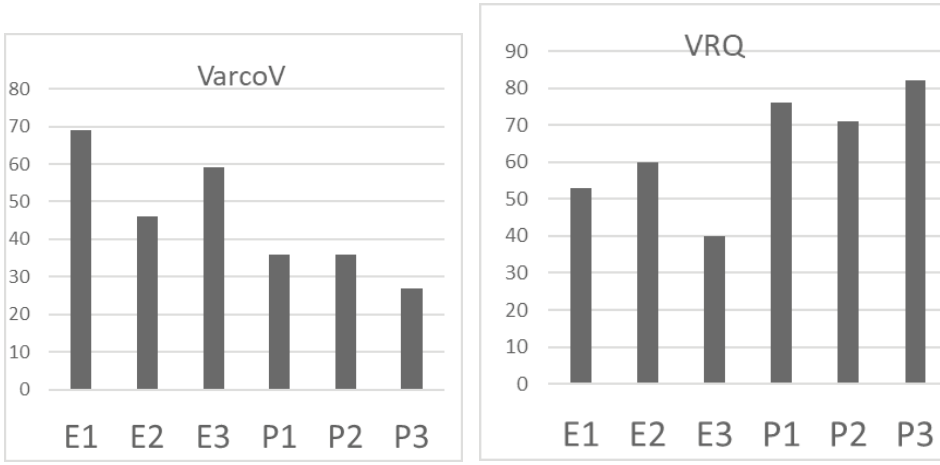

Figure 3. VarcoV and VRQ values for natural and backwards speech in *that the shop assistant was her daughter.*

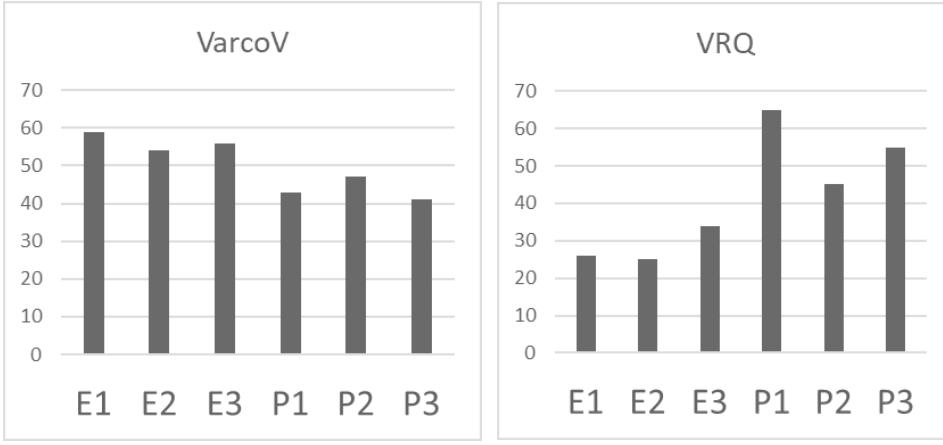

Figure 4. VarcoV and VRQ values for natural and backwards speech in *she chose one of the most expensive dresses in the shop.*

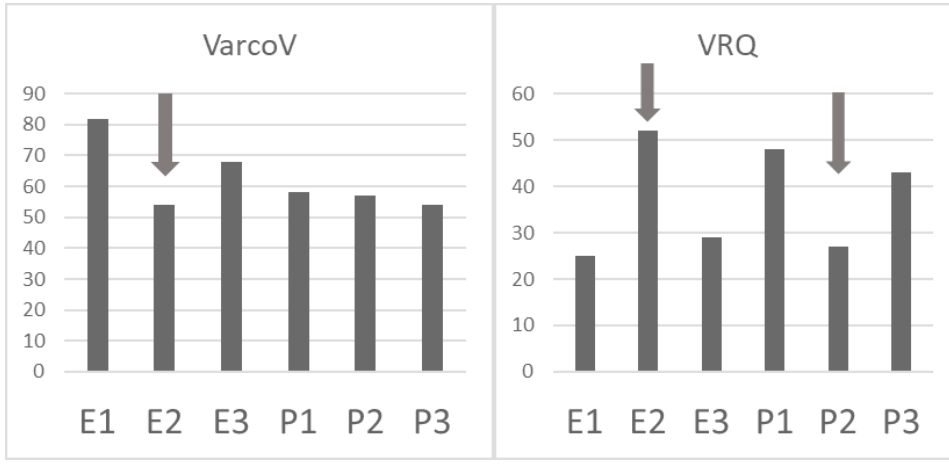

Figure 5. VarcoV and VRQ values for natural and backwards speech in *the temptation to steal is greater than ever before*. The values indicated by arrows are untypical of the group the speaker represents.

The untypical VarcoV and VRQ quotients observed in participants E2 and P2 performing phrase (T), indicated by vertical arrows, made these samples potentially interesting in the context of the study, by suggesting the opposite status of the speakers.

VarcoPeak has proved to be even more robust in separating Polish-accented from native English samples (Figure 6), with the exception of speaker E2 reading the phrase *that the shop assistant was her daughter.*

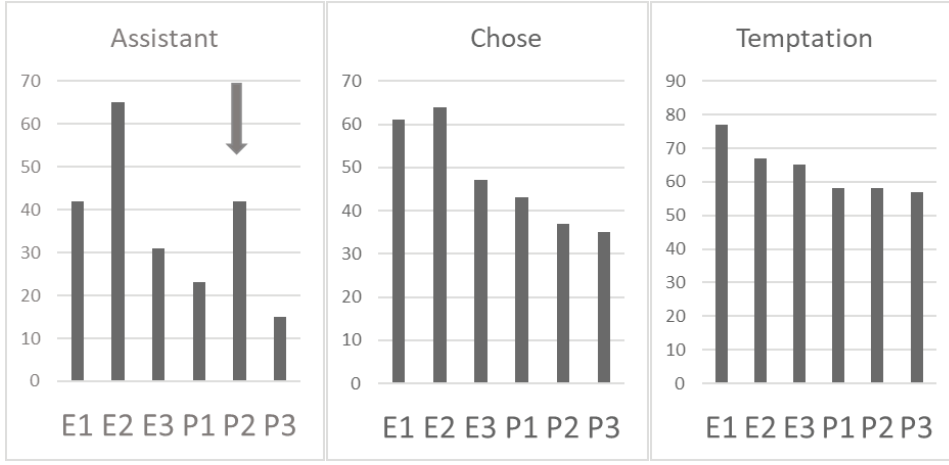

Figure 6. VarcoPeak values for vocoded speech in *that the shop assistant was her daughter* (Assistant), *she chose one of the most expensive dresses in the shop* (Chose), and *the temptation to steal is greater than ever before* (Temptation).

Procedure

The online experiment was designed using script in PsyToolkit (Stoet, 2010, 2017). The stimuli were blocked into vocoded, backwards, and natural. The blocks were presented in a fixed order so as to (1) expose the listeners to the most difficult task of identifying the vocoded speech at the beginning, and (2) to avoid carry-over influences from natural speech to vocoded and backwards speech. The stimuli in each block were randomized for each listener. The experiment started with collecting personal information, followed by three familiarization trials, each trial representing one of the three speech types. In every experimental trial, the listeners played a phrase by clicking on a 'play' button. The re-play option was not limited and the listeners were allowed to listen to each phrase as many times as they needed. All acoustic stimuli were accompanied by orthographic transcripts to allow access to the semantic content by parsing the acoustic information and thus facilitate the processing of temporal patterns (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005). After stimulus exposure, the listeners had to decide if they had heard a native speaker by clicking on 'yes' or 'no' buttons. They also had to indicate the certainty of their response on a 1–5 linear confidence scale by answering a question 'How sure are you?' from 1 'not at all' to 5 'very much.' Figure 7 shows the experiment interface.

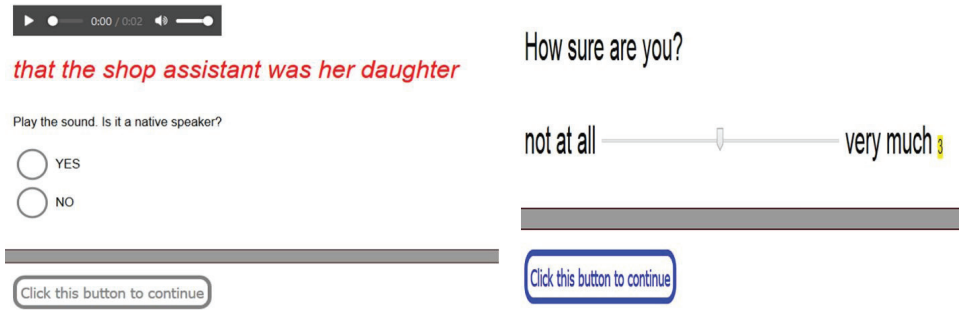

Figure 7. The experiment interface.

Altogether, there were 54 trials (6 speakers x 3 phrases x 3 speech types). The mean duration of all completed sessions was 18.5 minutes. The listeners were informed to use headphones or loudspeakers at a comfortable listening level.

Participants

Two groups of listeners were recruited to participate in the study. Sixteen native speakers of English, ten females and six males, who were recruited using social media. Fourteen of them reported British and three American accent. Another group were 14 advanced learners, 11 females and three males, recruited from the fifth year of the major English program at the Institute of English, University of Silesia in Katowice. Their proficiency in English ranged from C1 to C2 in CERFL, confirmed by regular curriculum tests. Out of the 30 recruited participants, only 14 completed the whole experiment. We will discuss the reasons for such a low completion rate in the Discussion section. We decided to include only the participants that had attended to all 54 trials. As a result, the analysis was carried out using the data from eight native speakers and six Polish advanced learners of English. The native speakers were five females and three males with the mean age of 36.6 years. British accent was indicated by six and American accent by two speakers. The advanced learners of English were five females and one male with the mean age of 24.1 years. Three of them claimed to be teachers of English. None of the listeners reported any speech or hearing disorders.

Analysis and Results

Listeners' responses were transformed to an *A'* sensitivity value (Donaldson 1992, 1993), which derives from the Signal Detection Theory (Green & Swets, 1966). It quantifies responses in terms of hits and false alarms. A value of 1 indicates perfect sensitivity and a value of 0.5 indicates performance at chance level. A value of 0 shows systematic confusion of all stimuli.

Initially, we planned to compare the performance of native speakers and advanced learners to see if there are significant differences between the two groups. However, due to the fact that only 14 listeners had completed all the trials, we ran independent-sample t-tests for each speech type to find if the results for the two groups might be pooled. The tests showed that the advanced learners did not perform differently from the native speakers in all three speech types: vocoded speech [$t(12) = 1.68$, $p = .12$], backwards speech [$t(12) = .62$, $p = .55$], natural speech [$t(12) = 1.46$, $p = .17$]. Consequently, we will analyse the data as collected from one group.

The Mixed Model ANOVA was designed with listener as a random effect and performance as a fixed effect (experimental level vs. chance level). The chance level was set at *A' M* = .05. The dependent variable was *A'* calculated for each listener and speech type. This model estimates the variance of random factors by constructing sums of squares and a cross products matrix for independent variables using Satterthwaite's method of denominator synthe-

sis (Luke, 2017; Searle, Casella, & McCulloch, 1992). The analyses revealed that the vocoded stimuli were not identified significantly above a chance level (M = .50; SE = .05) [F(1, 13) = .002, p = .96]. The mean confidence rating on a 1–5 scale with 1 'not certain at all' and 5 'very much certain' was 2.39 (SE = .09). The backwards-masked stimuli were not identified above chance level either (M = .59; SE = .05) [F(1, 13) = 4.03, p = .07]. The mean confidence rating was 2.85 (SE = .08). Predictably, natural speech was identified highly accurately (M = .96; SE = .01) [F(1, 13) = 2402.7, p <.001], with the mean confidence rating of 4.6 (SE = .04) (Figure 8).

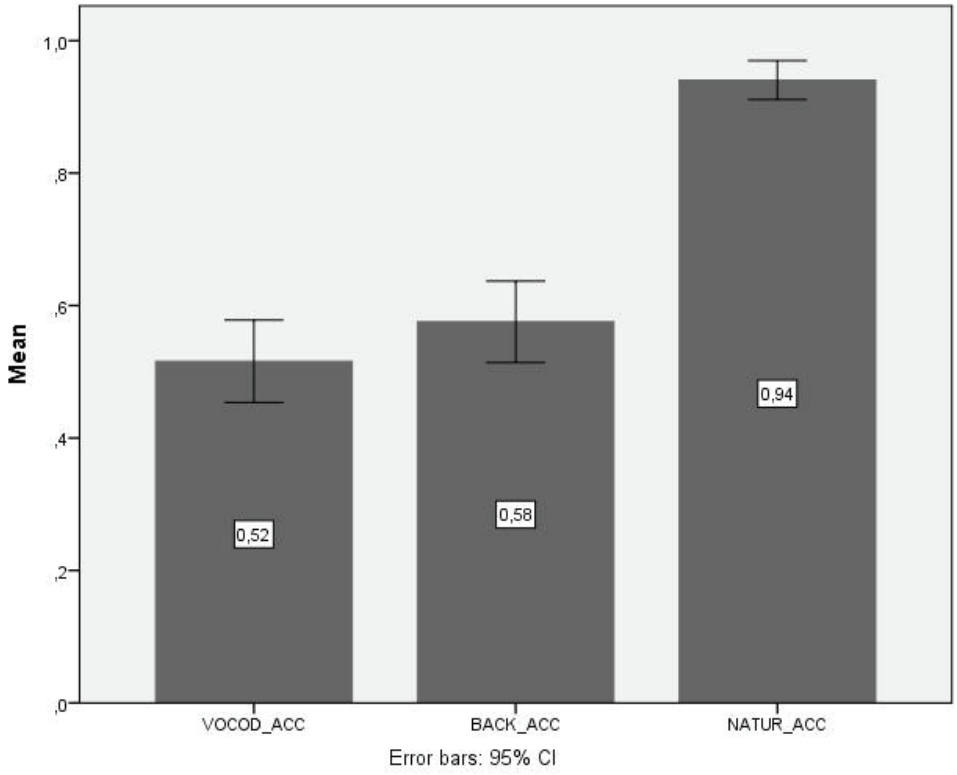

Figure 8. Mean *A'* scores for vocoded (VOCOD_ACC), backwards (BACK_ACC), and natural (NATUR_ACC) speech.

In order to explore more thoroughly the obtained results, we calculated the proportion of identification as native speaker for each speaker used in the experiment. The purpose was to investigate between-speaker variation to see if there were native speakers who were observably identified more frequently as non-native speakers and, by analogy, if there were Polish speakers who were more frequently identified as native speakers. We suspected that the listeners' global inability to identify the native/non-native status of the speakers may have

been contributed to by variation in individual speaker's scores. For the vocoded speech, the one-way ANOVA with speaker as an independent variable and proportion of identification as a native speaker as a dependent variable was not significant [$F(5, 12) = .77$, $p = .58$]. Figure 9 shows the values for each speaker.

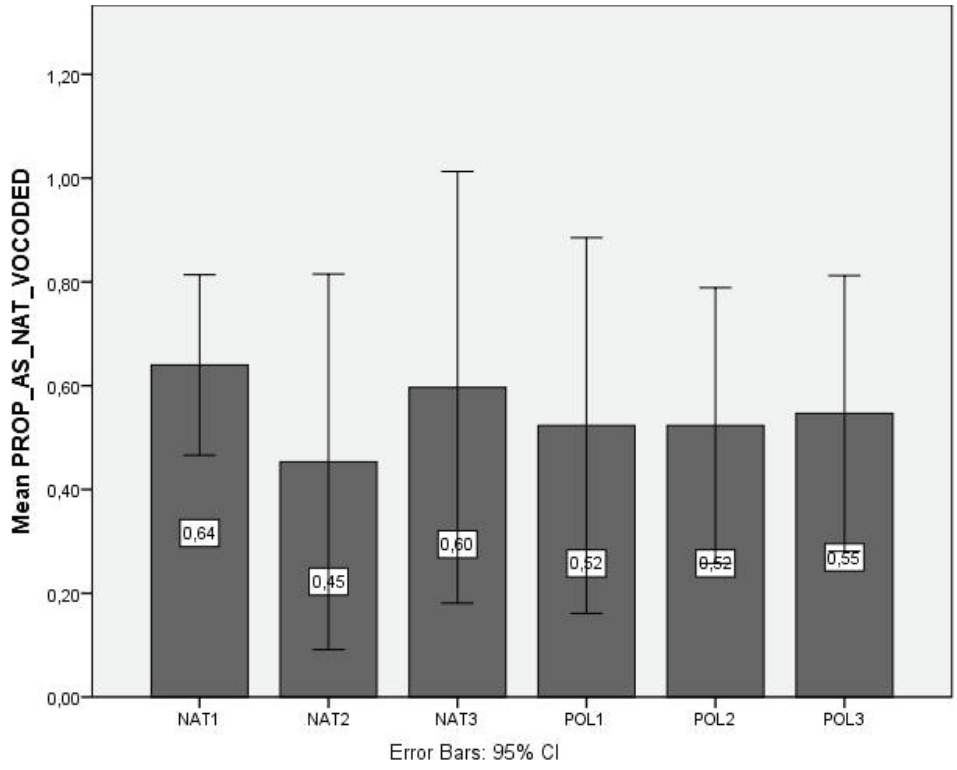

Figure 9. Proportion of identification as a native speaker in vocoded speech.

Although between-speaker variation was not significant, two patterns emerge from the data. Firstly, one native speaker (NAT 2) was identified as native less frequently than the other two. Secondly, error bars indicate relatively large deviations from the mean in all speakers except for NAT 1.

The same analysis for the backwards-masked speech revealed significant between-speaker variation [$F(5, 12) = 3.14$, $p = .049$]. Figure 10 presents the values for each speaker.

The results of this task show that two speakers stand out from the general confusion pattern. NAT 2 and POL 2 were more correctly identified as native and non-native speakers respectively. Strikingly, two other native speakers, NAT 1 and NAT 2, were reported to be native speakers only 43% and 45% of the time.

Finally, as discussed earlier, measures of temporal variability such as VarcoV, VRQ and VarcoPeak separated Polish-accented and native phrases

fairly robustly, so we ran Pearson correlations of those measures with the proportion of identification as a native speaker to find if they predicted the listeners' decisions in individual test phrases. Table 2 shows that none of these measures were significantly correlated with the listeners' performance in any of the three test phrases, indicating that the listeners were insensitive to temporal variation captured by these measures.

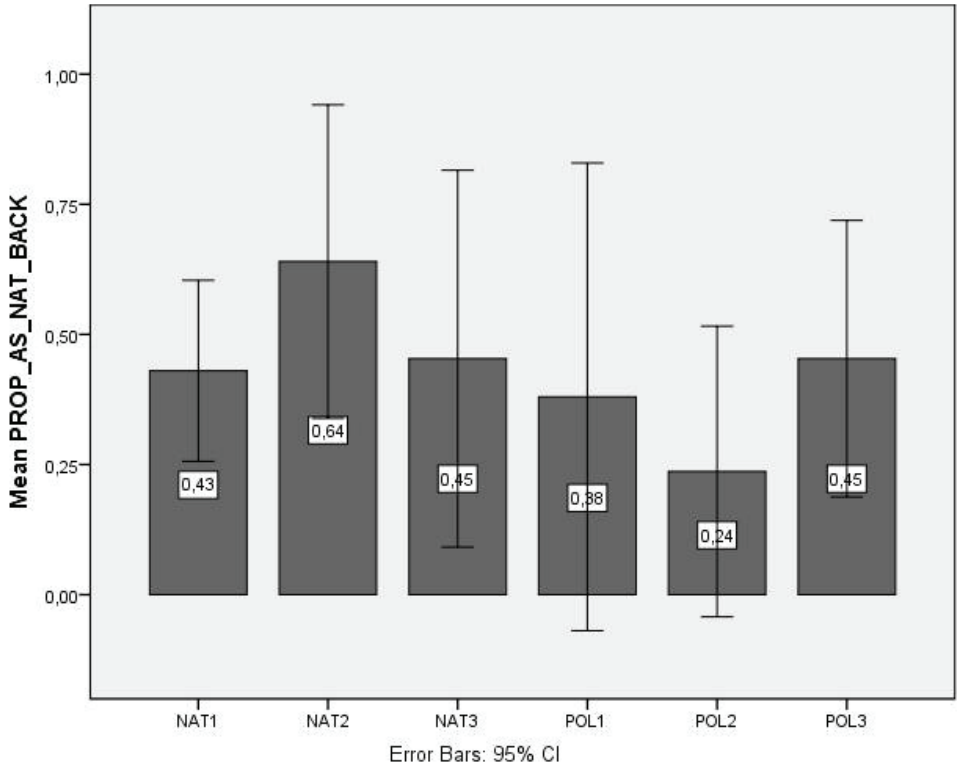

Figure 10. Proportion of identification as a native speaker in backwards-masked speech.

Table 2

Correlations of rhythm measures with the proportion of identification as a native speaker for each test phrase

	Backwards speech				Vocoded speech	
Phrase	VarcoV		VRQ		VarcoPeak	
	r	p	r	p	r	p
Assistant	–.09	.86	.17	.75	–.17	.74
Chose	.11	.83	–.14	.80	–.17	.74
Temptation	.04	.04	.09	.88	.35	.50

Discussion

In the present paper, we report findings that (a) Polish-accented and native English are not identified above-chance level from 4-band noise vocoded stimuli, (b) Polish-accented and native English are not identified above chance level from backwards-masked stimuli. In the following, we will relate the current results to the previously reported results.

Speech vocoding in accent recognition was used by Kolly and Dellwo (2014). They found that French-accented German and English-accented German could not be recognized above chance from 3-band vocoded speech samples and from 6-band vocoded samples without a speech transcript. However, above-chance performance was observed for 6-band vocoding with transcripts. It must be remembered that 3-band vocoding completely degrades the segmental make-up of speech, while 6-band vocoding may leave some spectral information about the quality of individual sounds. More recently, Kolly et al. (2017) showed that 6-band noise vocoded speech carries enough information for listeners to identify French- and English-accented German even when temporal cues are eliminated by means of duration transplantation, which further confirms that 6-band vocoding does not sufficiently degrade spectral cues. The contribution of the current study is that we used 4-band vocoding, which is not as degrading as 3-band vocoding, but is more effective in masking spectral cues than 6-band vocoding. The explanation for the current results that Polish-accented and native English were not identified successfully may be that 4-band vocoding is as degrading as 3-band vocoding and thus listeners are not able to separate Polish-accented and native English by syllable peaks only, without access to any spectral information.

Backwards-masking was used by Munro et al. (2010), who found that listeners distinguished native from non-native speech at above-chance levels with Mandarin, Cantonese, and Czech L2 speakers. The effect was robust enough to emerge from stimuli as short as one word as well as from randomly-spliced and monotone stimuli. The authors concluded that the listeners may have had access to the remnants of some sub-phonemic features or voice quality. In the current study, we provided the listeners with longer stimuli (more than seven words) and with authentic intonational contours and yet Polish speakers were not detected above chance. The difficulty with relating the current results with those in Munro et al. (2010) is the degree of accentedness that may be different between the tested groups. In the case of Mandarin and Cantonese speakers, it may be assumed that their accentedness was relatively high, because speakers of those languages are characterized by strong deviations from a native norm in segmental and prosodic realiza-

tions. Consequently, their successful detection may have been contributed to by their high degree of accentedness. On the other hand, Czech speakers, because of typological similarity of Polish and Czech, should be more comparable to Polish speakers; however, they were still successfully distinguished from native speakers with the mean *A'* score of .805.

A possible explanation for the differences between the current study and the previous studies is that this is the first study to test listeners outside a laboratory setting. The listeners in this study performed the task at home, using their own audio equipment, which means that the level of surrounding noise and specific parameters of the signal may have been different across listeners. This fact may have been responsible for the observed lower detection accuracy in this compared to the previous studies. However, it is not warranted to claim that the experimental conditions were not sufficiently optimal, because the mean *A'* score for natural speech was .94, which indicates that our listeners were able to detect non-native speakers highly effectively without content masking. Although laboratory conditions allow full control over the quality of stimulus delivery, it must be remembered that natural accent detection occurs in the real world, outside laboratory conditions, and still listeners perform highly effectively. If vocoded and backwards-masked speech provides sufficient cues to the speaker status, we should expect above-chance performance in and outside laboratory setting.

A methodological issue we want to raise is the level of engagement of listeners in tasks with content-masked stimuli. In this experiment, out of 30 recruited participants only 14 completed the whole session. We contacted some of them to ask for their motivation to quit before the experiment was finished. They reported that the blocks containing vocoded and backwards stimuli were confusing, irritating, and that, in their opinion, making a decision made no sense, because the signal contained no information about the speaker status. For this reason, although the statistics have been calculated on the basis of fourteen participants' responses, the information gathered from the remaining sixteen and the sheer fact of their withdrawal after some unsuccessful attempts to complete the task are clearly indicative of their inability to recognize the speaker's status, and in this way these respondents also contribute to the study and confirm its general results.

Although detection of speaker status from vocoded and backwards-masked speech was not overall successful, there was observable between-speaker variation in detection rate, especially in backwards speech. This may suggest that speakers may differ in robustness of individual acoustic cues that may or may not be masked. One of such cues, as suggested by Munro et al. (2010), may be voice quality, which is left intact in backwards speech. Future studies should attempt to directly correlate acoustic measures of speakers' global voice quality with accent detection.

Conclusions

The major limitation of the study is the fact that despite extensive research into speech rhythm, the notion still remains without a clear definition that would allow precise, objective description. The measures that we have employed, Varco and VRQ, may be criticized for not taking into account the non-temporal properties of rhythm, especially those related to prominence. On the other hand, they appear to be fairly efficient in separating non-native from native speech. Another problem is that any instrumental rhythm measures are sensitive to rather frequent timing deviation from prototypical values, which do not necessarily affect the listeners' accent judgements. To eliminate this type of variation, the experiment should include a large amount of stimuli, which in turn would make it even more strenuous for participants.

Bearing the limitations in mind, we argue that our study provides some new data concerning non-native accent detection. The results of the experiment show that neither native speakers nor Polish learners of English, who can easily recognize regular foreign-accented English speech samples, are able to detect native or non-native English accent from 4-band noise vocoded speech or backwards-masked speech. None of the tested rhythm measures of temporal variability, such as VarcoV, VRQ, and VarcoPeak, were correlated with the listeners' performance. Even though these rhythm measures are fairly robust for non-native accent detection, the speech characteristics that they refer to are not sufficient on their own for the listeners to identify the speaker's status. The general conclusion is thus that the rhythmic properties of speech alone, preserved in vocoded speech, or the temporal properties understood as syllable length variation together with voice quality, preserved in backwards speech, are not sufficient cues to foreign accent identification. This conclusion further implies that FL learning, which tends to focus on language detail more often than L2 acquisition, should not go too far in isolating individual aspects of pronunciation for classroom practice. Although focus on selected features of speech may be beneficial, the teacher must be aware that some of them may only perform their linguistic functions properly in interaction with other pronunciation components. The findings may therefore contribute to the debate on the approach to the relations between segments and prosody in foreign language teaching.

References

Alexander, L. G. (1967). *Practice and progress: An integrated course for pre-intermediate students*. London: Longman.

Andrianopolous, M. V., Darrow, K. N., & Chen, J. (2001). Multimodal standarization of voice among four multicultural populations: Formant structures. *Journal of Voice, 15*, 61–77.

Anisfeld, M., Bogo, N., & Lambert, W. E. (1962). Evaluational reactions to accented English speech. *Journal of Abnormal and Social Psychology, 65*, 223–231.

Arthur, B., Farrar, D., & Bradford, G. (1974). Evaluation reactions of college students to dialect differences in the English of Mexican-Americans. *Language Speech, 17*(3), 255–270.

Black, J. W. (1973). The 'phonemic' content of backward-reproduced speech. *Journal of Speech and Hearing Research, 16*, 165–174.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 10*, 341–345.

Cummings, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics, 26*, 145–171.

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.

Dellwo, V., Leemann, A., & Kolly, M.-J. (2012). Speaker idiosyncratic rhythmic features in the speech signal. *Electronic Proceedings of Interspeech 2012.* Portland, OR, USA, 1584–1587.

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics, 29*(3), 359–380.

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

Donaldson, W. (1993). Accuracy of d' and A' as estimates of sensitivity. *Bulletin of Psychonomic Society, 31*, 271–274.

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

Fourcin, A., & Dellwo, V. (2009). Rhythmic classification of languages based on voice timing. *UCL Eprints*. Retrieved from: http://eprints.ucl.ac.uk/15122/ accessed March 15, 2018.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Kolly, M.-J., Boula de Mareüil, P., Leemann, A., & Dellwo, V. (2017). Listeners use temporal information to identify French- and English-accented speech. *Speech Communication, 86*, 121–134.

Kolly, M.-J., & Dellwo, V. (2013). (How) can listeners identify the L1 in foreign-accented L2 speech? *Travaux Neuchâtelois de Linguistique, 59*, 127–148.

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.

Laver, J. (1980). *The phonetic description of voice quality.* Cambridge: Cambridge University Press.

Lee, C. S., & Todd, N. P. M. (2004). Towards an auditory account of speech rhythm: Application of a model of the auditory 'primal sketch' to two multi-language corpora. *Cognition, 93*, 225–254.

Lev-Ari, S., & Keysar, B. 2010. Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology, 46*, 1093–1096.

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

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods, 49*, 1494–1502.

Mennen, I. (2004). Bidirectional interference in the intonation of Dutch speakers of Greek. *Journal of Phonetics, 32*, 543–563.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition, 23*(4), 451–468.

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

Porzuczek, A. (2012). Measuring vowel duration variability in native English speakers and Polish learners. *Research in Language, 10*(2), 201–214.

Ramus, F., Hauser, M. D., Marc, D., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science, 288*, 349–351.

Raupach, M. (1980). Temporal variables in first and second language speech production. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech: Studies in honour of F. Goldman-Eisler* (pp. 263–270). The Hague: Mouton Publishers.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Process, 14*, 423–441.

Ryan, E. B., & Carranza, M. A. (1975). Evaluative reactions of adolescents toward speakers of standard English and Mexican American accented English. *Journal of Personality and Social Psychology, 31*(5), 855–863.

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

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.

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

Stoet, G. (2010). A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42*(4), 1096–1104.

Stoet, G. (2017). A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44*(1), 24–31.

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

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*, 628–639.

Toro, J. M., Trobalon, J. B., & Sebastián-Gallés, N. (2003). The use of prosodic cues in language discrimination tasks by rats. *Animal Cognition, 6*, 131–136.

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 Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters Part I: Recognition of backwards voices. *Journal of Phonetics, 13*, 19–38.

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

Arkadiusz Rojczyk, Andrzej Porzuczek

Erkennung eines fremden Akzents in der vokodierten und rückwärts gerichteten Sprache

Zusammenfassung

Die Studie befasst sich mit der Erkennung eines fremden Akzents in der englischen vokodierten und rückwärts gerichteten Sprache. Beide Verarbeitungsverfahren eliminieren eine semantische Information und teilweise (rückwärts gerichtete Sprache) oder vollständig (vokodierte Sprache) eine Spektralinformation, während die rhythmischen Merkmale der Sprache beibehalten werden, die als Differenzierungsgrad der Dauer von prosodischen Einheiten verstanden werden, die zur Unterscheidung von Proben des einheimischen und fremden Akzents dienen könnten. An der Untersuchung nahmen englische Muttersprachler und Polen teil, die diese Sprache auf fortgeschrittenem Niveau gebrauchen. Die Ergebnisse zeigten, dass weder Engländer noch Polen in der Lage sind, einen fremden Akzent in den verarbeiteten Sprachproben nur aufgrund der zeitlichen Verteilung der Akzente (vokodierte Sprache) und des Differenzierungsgrades der Länge von prosodischen Einheiten (rückwärts gerichtete Sprache) zu erkennen.

Schlüsselwörter: Akzenterkennung, nicht-muttersprachlicher Akzent, rückwärts gerichtete Sprache, vokodierte Sprache