Linguistica Silesiana 33, 2012
ISSN 0208-4228

ARKADIUSZ ROJCZYK
University of Silesia

# NON-NATIVE SPEECH PERCEPTION IN NOISE: THE VOICING CONTRAST IN ENGLISH

Speech is almost never delivered in ideal quiet conditions. On the contrary, the acoustic signal reaching a listener's ears is degraded by background noise and reverberations. The current study investigates the perception of the voicing contrast of initial stops in English by Polish non-native listeners. Previous research showed that Polish learners do not match native speakers of English in production and perception of English voiced and voiceless stops, which results from different phonetic implementations of voicing in the two languages. In the current study, two groups of Polish listeners recognised voicing of English initial stops in one-syllable words both in quiet and in six-talker babble. The results revealed different patterns of recognition for the two conditions. The place of articulation interacted significantly with voicing both in quiet and in noise, however results obtained suggest that performance in noise did not simply reflect the performance in quiet.

## 1. Introduction: speech in noise

In naturalistic conditions speech is very rarely delivered without the presence of background noise. Laboratory recordings obtained in quiet and further used in perception experiments do not fully reflect a complex nature of speech acoustics. Many landmark acoustic properties in the signal are actually masked by background noise and must be recovered by the listeners in order to sustain comprehension. It is the fact that talkers have been found to use clear speech, a speaking style used when talkers expect comprehension problems on the part of the listeners in noisy conditions, and that this speaking style increases intelligibility (Ferguson 2004; Payton *et al.* 1994; Picheny *et al.* 1985; Rogers *et al.* 2010; Uchanski *et al.* 1996). However, frequently talkers do not abandon their habitual speaking style even in degraded conditions. If this is the case, we may expect to observe a drastic decrease in comprehension, especially in groups of non-native listeners whose perceptual sensitivity is not fully developed. A similar decrease in comprehension can be observed in groups of older listeners whose perceptual sensitivity is already in decline (Dubno *et al*. 1984; Gordon-Salant and Yeni-Komshian 2010; Náběelek and Robinson 1982).

Experiments with masking by noise have been used for different purposes and with different types of maskers. Inclusion of background noise allows to tap the human auditory processes, which can be further used to e.g., calibrate noise-robust speech recognition systems (Hermansky 1990; Strope and Alwan 1997), or devise more efficient aids for the hearing impaired (Shannon *et al*. 1995). Types of maskers used most frequently in speech perception research include white noise, Gaussian noise, speech-shaped noise, single-talker maskers, or multi-talker babble with a set number of talkers (see Engen and Bradlow 2007). Although each of the maskers has its own acoustic properties that adversely influence recognition, the general observation is that intelligibility decreases as a function of growing similarity between a masker and the target voice (Brungart *et al*. 2001). In other words, talker-maskers will be more detrimental to performance in perception than noise maskers because they include human voice characteristics that will blend with the target voice. Listeners appear to be sensitive to the informational load of the talker-masker, as demonstrated by better speech intelligibility in time-reversed talker masker, the type of masker that retains voice characteristics but is semantically meaningless (Rheberhen *et al*. 2005), or babble modulated noise (Simpson and Cooke 2005). Moreover, target voice intelligibility has been found to decrease as additional voices are added into multi-talker babble (Bronkhorst 2000; Bronkhorst and Plomp 1992; Brungart *et al*. 2001; Rhebergen and Versfeld 2005; Simpson and Cooke 2005). Other research has shown that masking in multi-talker babble is more effective when the listeners know the language of the masker and when it is their native rather than second language (Garcia Lecumberri and Cooke 2006; Rhebergen *et al*. 2005). Finally, any type of the noise is manipulated using different signal-to-noise scales (SNR). SNR is defined as the ratio of signal power to the noise power corrupting the signal. Quite predictably, performance on speech recognition decreases with a decrease in SNR, which means that the level of noise increases relative to the level of target speech.

## 2. Non-native speech perception in noise

The sound categories in non-native language lack the perceptual stability compared to those of the native language. Comprehension in non-native language is rarely native-like even for proficient L2 speakers. What may not be so evident in optimal listening conditions, definitely manifests itself in suboptimal listening conditions. When listening to speech in noise, the perceptual system tries to make use of any acoustic cues currently available. The presence of background noise can force re-ranking of acoustic cues to linguistic categories, which results from the fact that primary cues are not available and the perceptual system seeks any secondary cues that can be extracted from the degraded signal (Jiang *et al*. 2006; Mattys *et al*. 2005; Van Engen and Bradlow 2007).

Lecumberri and Cooke (2006) reviewed studies that had compared native and non-native speech perception in noise and pointed to three basic findings that had emerged from those studies. First, native performance in noise was significantly better than non-native, even for early bilingual groups (Mayo *et al*. 1997). Second, increasing foreign-language proficiency decreases the negative effect of masking noise on perception (Florentine *et al*. 1984). Third, non-native listeners are not equally able to make use of linguistic context in decoding degraded speech relative to native speakers (Van Wijngaarden *et al*. 2004). Even specific differences within the groups of tested bilinguals have surfaced in experiments with perception in noise. Listeners with early L2 onset are characterised by better performance in noise that listeners with later L2 onset (Flege *et al*. 1999; Meador *et al*. 2000). Better performance has also been reported for non-native listeners with greater length of L2 exposure (Mayo *et al*. 1997). Other factors correlating positively with speech perception in noise are length of residence, amount of continued L1 use, and lexical structure (references in Pinet and Iverson 2010).

Differences in performance between native and non-native speakers are generally captured by two classes of explanations (Cutler *et al*. 2008). The first class locates those differences chiefly in identification on a phonemic level. Non-native speakers are at a disadvantage in that they use fewer acoustic cues for phonemic categories than native speakers. If their attention is directed to specific target cues exploited by native speakers, their performance in noisy condition improves (Hazan and Simpson 2000). Another class of explanations seeks to explain those differences at higher processing levels (Bradlow and Alexander 2007). Along this line of reasoning, auditory acuity of L1 and L2 listeners may be relatively equivalent, but the key difference lies in the effectiveness of recovery from degraded signal. Native speakers may be better at using all other than acoustic cues, such as, for example, phonotactic distributional patterns or semantic and contextual enhancements. This account has found support in studies reporting that non-native listeners are not always more adversely affected by degraded listening conditions (Bradlow and Bent 2002; Cutler *et al*. 2004; Cutler *et al*. 2008).

The extent to which native and non-listeners will differ in perception in noise is also determined by their ability to use higher-level processing to recover from disruption. In other words, native listeners may be more effective at semantic-contextual information at the sentence level (Bradlow and Alexander 2007). These results emerged in studies using high-probability and low-probability sentences embedded in noise (Benkí 2003; Cutler *et al*. 2004; Mayo *et al*. 1997).

## 3. The current study

In the current study we test the perception of the voicing contrast of initial stops in English by Polish learners. Differences in the implementation of the voicing contrast in English and Polish are best captured using the parameter of Voice Onset Time (Lisker and Abramson 1964). English voiced stops are produced with short-lag VOT values and voiceless stops are produced with long-lag VOT values. On the other hand, Polish uses voicing lead for voiced stops and short lag VOT for voiceless stops. This partial cross-category overlap results in the fact that English voiced stops have similar VOT to Polish voiceless stops. As a consequence, in production Polish learners implement insufficiently long VOT for English voiceless stops and the voicing lead for English voiced stops (Waniek-Klimczak 2005). In perception, Polish learners do not match native speakers in categorising the VOT continuum, in that they do not have a phonemic boundary along the positive VOT values (Rojczyk 2010), which is explained by the fact that Polish categorises voiced and voiceless stops at a 0 ms boundary (Mikoś *et al.* 1978). No previous study, to our knowledge, has tested directly the perception of English voiced and voiceless stops in quiet and in noise using Polish listeners.

Previous research demonstrated that perception of plosives in noisy environments is less robust than that of fricatives and that voicing is less affected by noise than the place of articulation (Miller and Nicely 1955). Jiang *et al.* (2006) concentrated more specifically on VOT in voicing perception in noise. They found that the voicing distinction was more difficult to perceive with decreasing SNR levels. Moreover, they reported that F1 onset frequency is more important for the perception of voicing at low SNRs. It results from the fact that VOT duration is less immune to noise masking than formant frequencies.

### 3.1. Participants

A total of 35 listeners participated in the study. Gender was not balanced – there were 29 females and 6 males. They were recruited from second-year students at the Institute of English, University of Silesia. Such a selection guaranteed a fairly uniform level of proficiency. They all considered themselves as advanced speakers of English with no difficulties in communicating with native speakers. Their age ranged from 20 to 24 (*M* = 20.3). All participants received a partial course credit for their participation. None of the subjects reported any hearing disorders nor had any history of such.

The listeners were quasi randomly ascribed to two groups. Seventeen listeners listened to the stimuli in silence and 18 listeners were presented with the same stimuli in noise.

### 3.2. Stimuli

Eighteen monosyllabic words were created that included three voiced and three voiceless plosives for each place of articulation (bilabial, alveolar, velar). All 18 stimuli (2 voicing categories x 3 places of articulation x 3 vowel contexts) are presented in Appendix A. The voiced-voiceless oppositions were obtained by providing the same VC context (e.g., *tan* vs. *dan* or *cot* vs. *got*). The word list was presented to a qualified phonetician who recorded them using a British pronunciation model in a carrier phrase *I said ... of course*. The reader was instructed to produce phrases using neutral falling intonation. The recording took place in the Acoustics Laboratory at the Institute of English, University of Silesia, in a sound-proof booth. The signal was captured with a headset condenser microphone Sennheiser HME 26-600S, preamplified with USBPre 2 (Sound Devices) into .wav format with the sampling rate 48KHz and 24 bit quantization.

VOT in each stimulus was measured between the first peak of the release burst to the onset of the second formant of the following vowel using Praat 5.1.38 (Boersma 2001). All stimuli were characterised by VOT values that neatly separated voiced and voiceless stops into short lag and long lag categories (Appendix A). None of the voiced stops had negative VOT values.

All 18 recorded phrases were subsequently peak normalized for amplitude at 70 dB and saved as individual .wav file for a perception experiment.

A noise component was obtained from English six-talker babble created by merging 6 talkers into one sound file. The amplitude was scaled to 70dB. Such noise was added to target phrases at 0 SNR exceeding its duration at the onset and offset by 500 ms. Next, Cool Edit Pro 7.9.9. was used to fade in and out noise in each phrase. Such a technique was deemed to provide the most comfortable introduction of the noise component into target recordings.

### 3.3. Procedure

The experiment took place in a quiet room at the Institute of English, University of Silesia. The stimuli were presented binaurally through Philips SBC HP840 headphones at a comfortable listening level. Each stimulus was presented four times, which gave 72 trials (18 stimuli x 4 repetitions) and the presentation order was randomized for each listener. A response sheet was provided with 72 carrier phrases *I said ... of course* in which the target word was replaced by a dotted line. The listeners were required to write down the whole word they had heard. Each experimental session lasted approximately 20 minutes and was preceded by a short conversation in English during which the experimenter provided instructions and answered questions if necessary. The listeners participating in the noise session were additionally informed that they would hear sentences embedded in noise and that they should do their best to identify the missing words.

The responses were analysed using the voicing criterion as a primary reference. It means that responses that were incorrect for the place of articulation (e.g., *sap* instead of *tap*) were counted as correct as long as the voicing category was the same (in this case – voiceless). All responses that had the wrong place of articulation were, however, identified and included for further analysis. The representation of vowels was not analysed.

## 4. Results

A two-way factorial ANOVA was designed with 2 (voiced/voiceless) x 3 (bilabial/alveolar/velar) independent variables and a correct recognition rate as a dependent variable. Statistics was calculated separately for each of the experimental conditions: perception in silence vs. perception in noise.

### 4.1. Perception in quiet

The analysis revealed a significant main effect of voicing in perception in quiet [F(1, 304) = 26.758, p < .001], indicating that voiced stops were significantly more difficult to correctly recognise than voiceless stops. The main effect of place of articulation was also significant [F(2, 303) = 16.282, p < .001], revealing a regular pattern of bilabial stops obtaining the highest recognition scores and velar stops the lowest recognition scores. Post Hoc Fisher LSD tests indicated that the recognition of bilabial and alveolar stops did not contribute significantly to the main effect (p = .41). In other words, velar stops were much more difficult to perceive relative to bilabial and alveolar stops, but there was no significant difference between the latter two. The interaction analysis showed a significant interaction of the place of articulation and voicing on the correct recognition rate [F(2, 300) = 19.617, p < .001]. The Post Hoc analysis determined that this effect was mainly caused by a markedly low recognition of voiceless velar stops, as demonstrated in Figure 1.

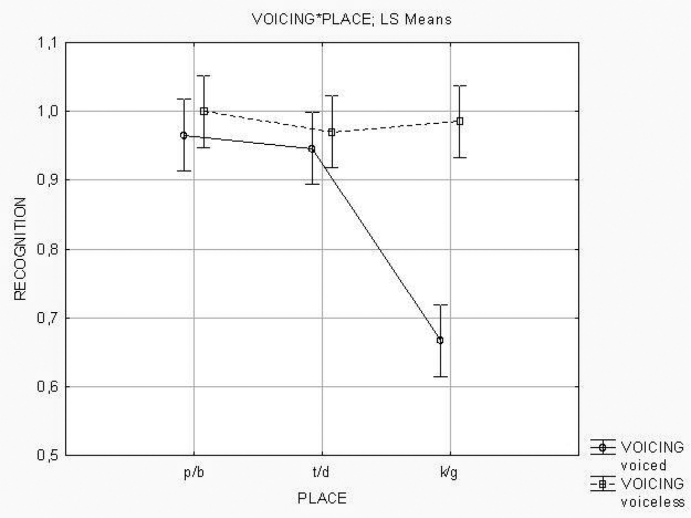

Fig 1: Interaction between the place of articulation and voicing on recognition rate in silence

### 4.2. Perception in noise

In the noise condition, voicing did not have a significant main effect on correct recognition [F(1, 322) = .127, p = .72]. The lack of the main voicing effect was caused by lower perception scores for voiceless stops in noise relative to those in silence. The main effect of the place of articulation was significant [F(2, 321) = 4.202, p = .05]. Post Hoc Fisher LSD test indicated that alveolar /t, d/ had a significantly higher recognition rate than both bilabial /p, b/ (p < .05) and velar /k, g/ (p < .01). There was no individual difference between the latter two. There was a significant interaction of the place of articulation and voicing on the correct recognition [F(2, 318) = 20.312, p < .001]. As determined by the Post Hoc analysis, this interaction effect was contributed to by lower recognition scores for voiceless bilabial /p/ and voiced velar /g/ (Figure 2).

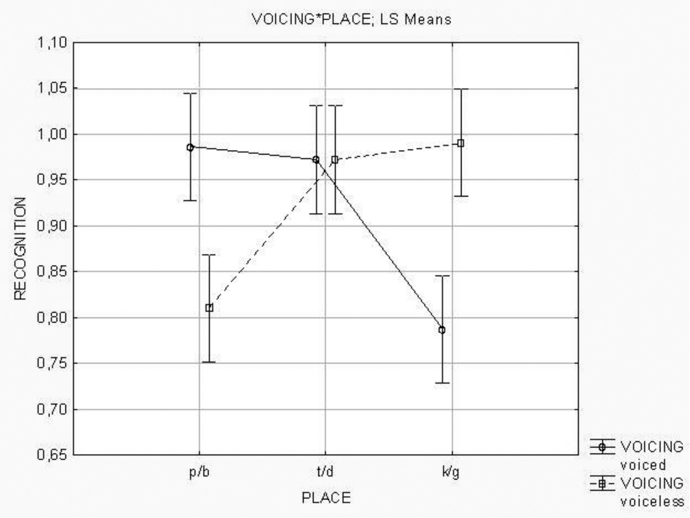

Fig 2: Interaction between the place of articulation and voicing on recognition rate in noise.

## 5. General discussion

The purpose of the study was to analyse the perception of English word-initial voiced and voiceless plosives in noise by Polish learners. To this end, two groups of listeners listened to the one-syllable stimuli in silence and in six-talker babble. It was hypothesised that the degrading influence of noise would allow to determine actual perception performance in the tested group. Speech is very rarely delivered in ideal conditions. Perception in noise is believed to be a better determinant of a non-native listener's level of sound acquisition, because it taps more directly into competence (Garcia Lecumberri and Cooke 2006).

The results revealed that in silence voiceless stops are better recognised than voiced stops. These results are in congruence with earlier studies (Rojczyk 2010) that demonstrated a category overlap between Polish voiceless short lag and English voiced short lag stops. However, in noise this difference was eliminated. The perception rate of both voiced and voiceless stops leveled, which was mainly achieved by lower recognition rates of voiceless stops in noise relative to in silence. Counter to our hypothesis, the magnitude of noise masking was not larger for voiced stops in noise. While the recognition rate for voiced stops was similar in the two conditions, voiceless stops suffered more from degraded conditions.

The place of articulation had a significant effect on perception both in quiet and noise, however the observed pattern differed dramatically in the two conditions. Moreover, the place of articulation interacted significantly with voicing. In silence, bilabial and alveolar stops were perceptually more robust than velar stops. However, the actual interaction between the place of articulation and voicing in Figure 1 shows that less effective perception of velar stops was strongly motivated by perception rates of voiced velars. In other words, listeners' performance was dramatically lower for voiced velars compared to other categories.

The interaction analysis of the place of articulation and voicing in noise also indicated a significantly lower perception rate of voiced /g/ (Figure 2). Moreover, unlike in silence, voiceless /p/ had a similarly lower recognition rate. The alveolar place of articulation was characterised by stable rates irrespective of the voicing contrast.

The results reported here seem to suggest that results obtained from perception experiments in silence may not always provide a complete picture of actual performance of non-native listeners. First of all, counter to predictions voiced stops were not more difficult to recognize in noise for Polish non-native listeners. On the contrary, perception of voiceless stops was more degraded. Why this may be the case is hard to explain, considering the fact that long lag VOT values should serve as a strengthening perceptual cue. At least, it is suggested by results from perception in quiet. Secondly, the interaction between the place of articulation and voicing demonstrated an inconsistent pattern both in quiet and in noise. What they had in common though was the relative high perception rate of alveolars and low perception rate of voiced velars.

More research on non-native speech perception with Polish listeners should include noise masking as one of the conditions. The experiment reported here suggests that research on perception in noise may provide results that will observably differ from the ones obtained from perception in quiet. Future studies will have to systematise them and suggest more comprehensive explanations.

## Appendix A

Words used in the experiment with measurements of VOT in miliseconds.

| Voiceless | VOT | Voiced | VOT |
|---|---|---|---|
| pat | 95 | bat | 17 |
| Pete | 73 | beat | 19 |
| pet | 64 | bet | 16 |
| tan | 121 | dan | 31 |
| teen | 99 | dean | 42 |
| ten | 87 | den | 38 |
| cap | 94 | gap | 34 |
| cot | 82 | got | 37 |
| cut | 58 | gut | 26 |

**Acknowledgements**
This work has been partly supported by a grant from Polish Ministry of Science and Higher Education (0576/B/H03/2010/38). We thank Andrzej Porzuczek for recording stimuli for the study.

## References

Benkí, J. R. 2003. Quantitative evaluation of lexical status, word frequency, and neighbourhood density as context effects in spoken word recognition. *Journal of the Acoustical Society of America* 113: 1689-1705.

Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 10: 341-345.

Bradlow, A. R., and J. A. Alexander. 2007. Semantic-contextual and acoustic-phonetic enhancements for English sentence-in-noise recognition by native and non-native listeners. *Journal of the Acoustical Society of America* 117: 2339-2349.

Bradlow, A. R., and T. Bent. 2002. The clear speech effect for non-native listeners. *Journal of the Acoustical Society of America* 112: 272-284.

Bronkhorst, A. W. 2000. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acoustica* 86: 117-128.

Bronkhorst, A. W., and R. Plomp. 1992. Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *Journal of the Acoustical Society of America* 92: 3132-3138.

Brungart, D. S., B. D. Simpson, M. A. Ericson and K. R. Scott. 2001. Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Journal of the Acoustical Society of America* 110: 2527-2538.

Cutler, A., M. Garcia Lecumberri and M. Cooke. 2008. Consonant identification in noise by native and non-native listeners: Effects of local context. *Journal of the Acoustical Society of America* 124: 1264-1268.

Cutler, A., A. Weber, R. Smits and N. Cooper. 2004. Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America* 116: 3668-3678.

Dubno, J. R., D. D. Dirks and D. E. Morgan. 1984. Effects of age and mild hearing loss on speech recognition. *Journal of the Acoustical Society of America* 76: 87-96.

Ferguson, S. H. 2004. Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *Journal of the Acoustical Society of America* 116: 2365-2373.

Flege, J. E., I. R. A. MacKay and D. Meador. 1999. Native Italian speakers' perception and production of English vowels. *Journal of the Acoustical Society of America* 106: 2973-2987.

Florentine, M., S. Buus, B. Scharf and G. Canevet. 1984. Speech reception thresholds in noise for native and non-native listeners. *Journal of the Acoustical Society of America* 75: 84.

Garcia Lecumberri, M. L., and M. Cooke. 2006. Effect of masker type on native and non-native consonant perception in noise. *Journal of the Acoustical Society of America* 119: 2445-2454.

Gordon-Salant, S., and G. H. Yeni-Komshian. 2010. Recognition of accented English in quiet and noise by younger and older listeners. *Journal of the Acoustical society of America* 128: 3152-3160.

Hazan, V., and A. Simpson. 2000. The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects. *Language and Speech* 43: 273-294.

Hermansky, H. 1990. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America* 33: 589-596.

Jiang, J., M. Chen and A. Alwan. 2006. On the perception of voicing in syllable-initial plosives in noise. *Journal of the Acoustical Society of America* 119: 1092-1105.

Lisker, L., and A. S. Abramson. 1964. A cross language study of voicing in initial stops: Acoustic measurements. *Word* 20: 384-422.

Mattys, S. L., L. White and J. F. Melhora. 2005. Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology* 134: 477-500.

Mayo, L. H., M. Florentine and S. Buus. 1997. Age of second-language acquisition and perception of speech in noise. *Journal of Speech, Language, and Hearing Research* 40: 686-693.

Mikoś, M. J., P. A. Keating and B. J. Moslin. 1978. The perception of voice onset time in Polish. *Journal of the Acoustical Society of America* S1: 63.

Miller, G. A., and P. E. Nicely. 1955. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America* 27: 338-352.

Nábělek, A. K., and P. K. Robinson. 1982. Monaural and binaural speech perception in reverberation for listeners of various ages. *Journal of the Acoustical Society of America* 71: 1242-1248.

Payton, K. L., R. M. Uchanski and L. D. Braida. 1994. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *Journal of the Acoustical Society of America* 95: 1581-1592.

Picheny, M. A., N. I. Durlach and L. D. Braida. 1985. Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research* 28: 96-103.

Pinet, M., and P. Iverson. 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: 1357-1365.

Rhebergen, K. S., and N. J. Versfeld. 2005. A Speech-Intelligibility Index- based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *Journal of the Acoustical Society of America* 114: 2181-2192.

Rhebergen, K. S., N. J. Versfeld and W. A. Dreschler. 2005. Release from informational masking by time reversal of native and non-native interfering speech. *Journal of the Acoustical Society of America* 118: 1274-1277.

Rogers, C. L., T. M. DeMasi and J. C. Krause. 2010. Conversational and clear speech intelligibility of /bVd/ syllables produced by native and non-native English speakers. *Journal of the Acoustical Society of America* 128: 410-423.

Rojczyk, A. 2010. *Temporal and spectral parameters in perception of the voicing contrast in English and Polish.* Katowice: Wydawnictwo Uniwersytetu Śląskiego.

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

Simpson, S. A., and M. Cooke. 2005. Consonant identification in N-talker babble is a non-monotonic function of N. *Journal of the Acoustical Society of America* 118: 2775-2778.

Strope, B., and A. Alwan. 1997. A model of dynamic auditory perception and its application to robust word recognition. *IEEE Transactions on Speech and Audio Processing* 5: 451-464.

Uchanski, R. M., S. S. Choi, L. D. Braida, C. M. Reed and N. I. Durlach. 1996. Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *Journal of Speech and Hearing Research* 39: 949-509.

Van Engen, K. J., and A. R. Bradlow. 2007. Sentence recognition in native- and foreign-language multi-talker background noise. *Journal of the Acoustical Society of America* 121: 519-526.

Van Wijngaarden, S. J., H. J. M. Steeneken and T. Houtgast. 2002. Quantifying the intelligibility of speech in noise for non-native listeners. *Journal of the Acoustical Society of America* 111: 1906-1916.

Waniek-Klimczak, E. 2005. *Temporal parameters in second language speech: An applied linguistic phonetic approach*. Łódź: Wydawnictwo Uniwersytetu Łódźkiego.