Title: We do not perceive just speech sounds, we perceive individual speakers' characteristics : indexical properties in speech processing

Author: Arkadiusz Rojczyk

Citation style: Rojczyk Arkadiusz. (2009). We do not perceive just speech sounds, we perceive individual speakers' characteristics : indexical properties in speech processing. W: H. Fontański, R. Molencki, O. Wolińska, A. Kijak (red.), "W kręgu teorii : studia językoznawcze dedykowane profesorowi Kazimierzowi Polańskiemu in memoriam" (S. 185-194). Katowice : Wydawnictwo Uniwersytetu Śląskiego.

Arkadiusz Rojczyk

We do not perceive just speech sounds, we perceive individual speaker's characteristics: indexical properties in speech processing

The traditional approach to speech perception has relied on the assumption that speech is structured in systematic ways and that the linguistic information encoded in the speech signal can be represented reliably and economically as a sequence of abstract, linear units. Speech has been thought to be "basically a sequence of discrete elements" (Licklider 1952: 590), for "in writing we perform the kind of symbolization (…) while in reading aloud we execute the inverse of this operation: that is, we go from a discrete symbolization to a continuous acoustic signal" (Halle 1956: 510). The word *feel* is traditionally represented as composed of three segments /f/ /i:/ /l/, sequenced in a linear fashion. It is differentiated from the word *veal* /vi:l/ by the feature of voicing, with /f/ being voiceless and /v/ being voiced. Segmental representations are thus designed to code only the linguistically significant differences in meaning between minimal pairs of words in the language (Twaddell 1952).

1. Speech is not a sequence of invariant, discrete units

If, as claimed by traditional theorists, human perceptual system operates on sequences of discrete, linear segments, certain crucial assumptions must be made (discussion in Pisoni and Levi 2007). First, continuous and gradient information in the speech signal can be represented by a set of discrete and linear symbols (Pierrehumbert and Pierrehumbert 1990). Second, units applied to speech representation must be abstract and context-free. Third, speech perception relies on psychological mechanisms that normalize acoustically different speech signals to make them functionally equivalent in perception (Joos 1948).

None of these assumptions stands empirical validation. The acoustic signal is not linear because coarticulation and other contextual effects necessarily destroy linearity and discreteness. For example, perceptual information about the place of articulation for /b/, /d/, /g/ are located in formant transitions for a following vowel. Therefore, the period of vowel onset contains both information about preceding stops and ongoing vowels (Liberman et al. 1967). The acoustic signal does not contain easily separable units. Rather, different speech sounds penetrate one another spreading their acoustic identity on other segments (Hockett 1955). Finally, speech sounds encoded in the signal are not invariant. Invariance entails that every phoneme must have a specific set of acoustic attributes in all contexts (Murphy 2002). This is definitely not the case. Significant across-speaker variation results in different acoustic realisations of particular sounds for men, women, and children due to differences in vocal tract length. A classic study by Peterson and Barney (1952) demonstrated that one's person *bet* might be another person's *bit*. Different vocal tract anatomy between men and women results in highly variable formant values for male and female vowels (Johnson 2008). Consider, for example, the following Figures 1 and 2 representing the productions of a Polish word *beż* (beige) by a male and female speaker. All three formants for vowel /e/ extracted from a steady-state portion differ in absolute values across gender. Male values are 584 Hz for Formant 1, 1583 Hz for Formant 2, and 2515 Hz for Formant 3. Female values are 678 Hz for Formant 1, 1890 Hz for Formant 2, and 2846 for Formant 3.

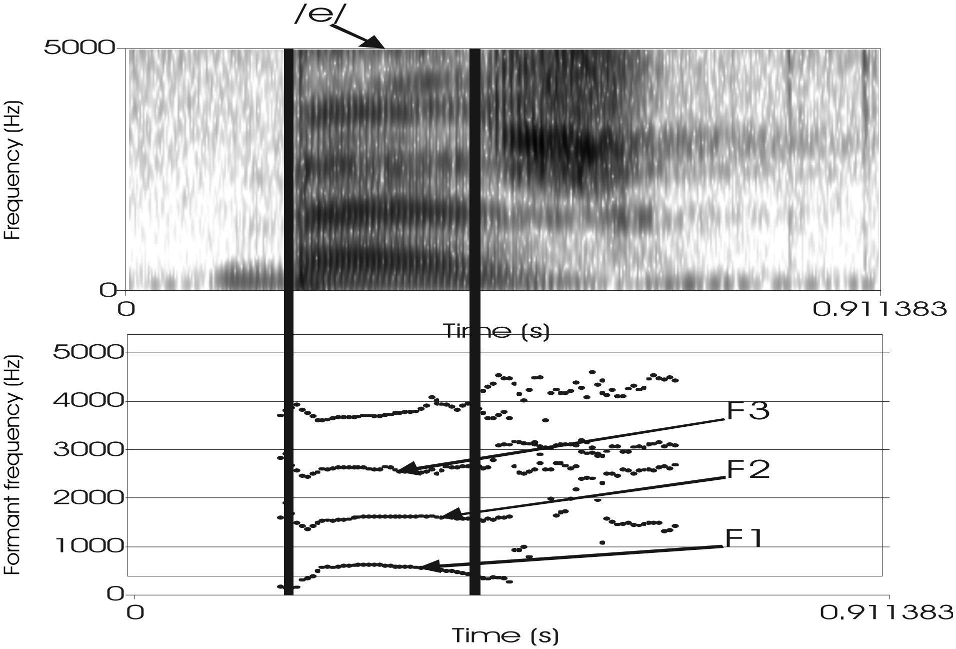

Figure 1. Spectrogram and formant tracking display for male voice. Polish word *beż* (beige)

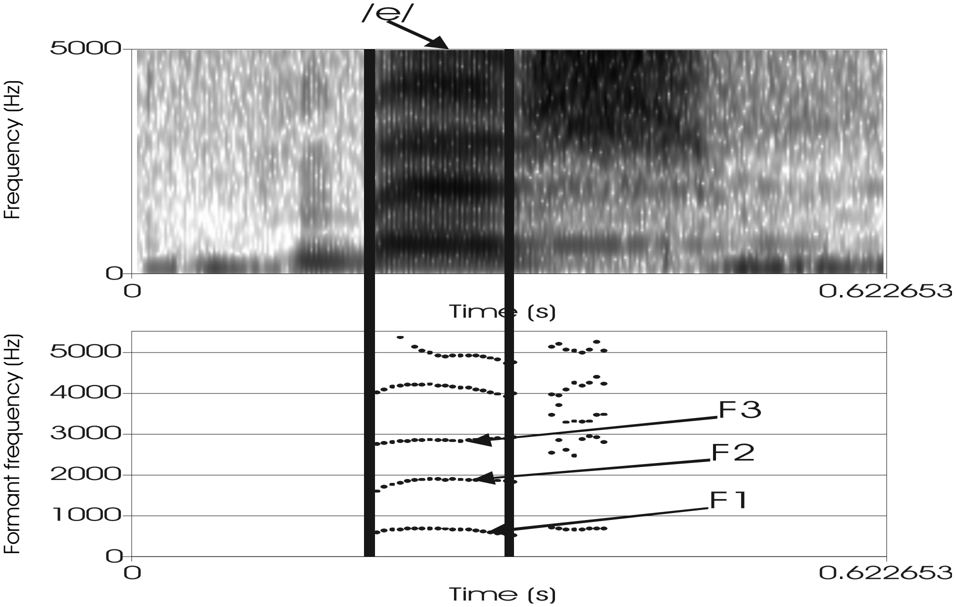

Figure 2. Spectrogram and formant tracking display for female voice. Polish word *beż* (beige)

Multiple talker-specific variability sources have been documented in the literature so far. These are not only anatomical differences across speakers (Bradlow et al. 1996) but also articulation rate (Miller et al. 1984), articulation effort (Moon and Lindblom 1994), speaking style (Uchanski 2008), or even discourse elements such as turn-taking rules (Local 2002 reported in Nygaard 2008). Talker's idiosyncratic characteristics conveys information about their identity, emotional state, region of origin, or health and age (review in Clopper and Pisoni 2008, Nygaard 2008, Vaissiére 2008).

The traditional approach assumed that any talker-specific variability is discarded during a recognition process so that only abstract units of speech are extracted from the speech signal. Any additional variation, as reviewed above, was claimed to be processed separately from a segmental level. In other words, features such as talker identity, emotional tone of voice, accent or dialect were considered to be non-linguistic and thus excluded from the recovery of linguistic content. The following sections of the article demonstrate a body of research which casts doubt on the traditional separation between linguistic (segmental) and nonlinguistic (talker-specific) information in the speech signal[1]. Results from psycho-acoustic experiments point to the conclusion that surface form is not discarded during the processing of spoken language but rather it is retained and used. The episodic approaches considered here (Goldinger 1998, Nygaard 2008, Pisoni and Levi 2007) assume that spoken words are represented in lexical memory as a collection of individual perceptual tokens rather than as a sequence of abstract segments. Along this line of reasoning, listeners encode specific instances of perceptual episodes. Simplifying it considerably for the purpose of clarity, we may say that any speaker of Polish stores in long-term memory all the instances of a word *pisać* he or she has ever heard, retaining talker-specific details of particular talkers.

[1] The reviewer pointed to the fact that the research referred to below, which is mostly seated in acoustic phonetics, does not invalidate the tenets of phonological description and its concept of a phoneme. Obviously, it was not our aim to question the applicability of phonological abstract concepts in the description of languages and their sound pattern both synchronically and diachronically. Rather, we wanted to demonstrate, by referring to psycholinguistc experiments, that speech processing cannot be completely understood if a classical parsing of a signal into phonemes is claimed to be the only medium of perception.

2. The influence of individual speaker's characteristics on speech processing

Several experiments carried out in the Speech Research Laboratory at Indiana University have demonstrated that indexical properties of a talker are not only integrated early in the process of speech perception but are also stored in long-term memory for further retrieval (Pisoni and Levi 2007). Mullenix and colleagues (1989) observed that intelligibility of spoken words presented in interfering and distracting noise depended on the number of voices used in stimulus presentation. In one condition, all the test words were produced by a single talker, in another condition, fifteen different talkers produced the same test words. The authors presented subjects with the test words in three different signal-to-noise ratios. Identification performance of the test stimuli was always better when the words were produced by a single talker rather than by multiple talkers. The results point to the conclusion that talker-specific characteristics influences the processing and recognition of speech.

The claim that linguistic information (speech sounds) is processed separately from extralinguistic properties (indexical, talker-specific properties) has been undermined in a study by Mullenix and Pisoni (1990). They asked subjects to attend selectively either to talker's voice or phoneme identity in presented test words. Thus in one condition, the subjects were asked to concentrate on the characteristics of talker's voice and simultaneously ignore phonemic structure of the stimuli. In another condition, the subjects were required to concentrate on the meaning and ignore the indexical properties of voices. The authors concluded that words and voices are not processed separately because otherwise interference from non-attended dimension should not have been observed.

Not only are indexical properties concurrently processed in speech recognition, but they are also stored in long-term memory. Goldinger and colleagues (1991) found that attributes of talker's voice persist and participate in speech recognition up to a week after perceptual analysis has been completed. The very process of learning speaker's voice characteristics appears to be an indispensable component of effective meaning extraction from the speech signal. Nygaard et al. (1994) trained listeners to learn to identify a set of ten talkers' voices from single word utterances over a period of nine days. Every day listeners were familiarised with a single talker and learned to associate a common name with each talker's voice. After the learning period, subjects were asked to recognise a set of novel words (not used in the training phase) with increasing signal-to-noise ratios. One group of listeners transcribed words produced by the talkers that they had been trained on. Another group of listeners transcribed words produced by new talkers with whose voices they had not been familiarised. The results clearly revealed that subjects who heard novel words produced by familiar talkers had much more accurate recognition than subjects who heard unfamiliar voices. The authors conclude that we may speak of a process of perceptual adaptation or rapid tuning that occurs during processing speech from different talkers.

More recent results suggest that the processing of indexical properties of individual speakers may be independent from linguistic content. Winters et al. (2006 reported in Pisoni and Levi 2007) trained two groups of monolingual English subjects to identify ten voices speaking either in English or German. After a training period of four days, listeners were asked to recognise the same ten voices but in an untrained language. Accordingly, subjects trained on voices speaking English were now asked to classify them speaking German and subjects familiarised with voices speaking German were now to identify them speaking English. The results demonstrated that listeners from each group were able to generalize stored properties of a particular voice to the untrained language thus being independent from linguistic content.

Since a speaking rate is one of individual, idiosyncratic, properties of each speaker, it is of no surprise that it is effectively encoded along other voice parameters. Bradlow and Pisoni (1999) presented listeners with lists of words produced at different speaking rates: slow, medium, fast by twenty different voices (ten male and ten female). The list contained words with varying difficulty. Easy words were high frequency words and difficult words were low frequency words. The authors had subjects transcribe words produced by each talker at each speaking rate. The results revealed that subjects learnt particular talker's speech rate characteristics and that this experience improved recognition for both difficult and easy words. Interestingly enough, transcription performance as a function of experience with a particular talker speech rate improved more for difficult than for easy words.

The importance of individual speaker's indexical properties encoded in his or her speech has been additionally demonstrated in second language acquisition literature. Consistent with the traditional approach that any speaker-specific properties are discarded in speech processing and that only discrete, contrastive segments are extracted from the signal, Strange and Dittmann (1984) attempted to teach adult Japanese learners a phonetic contrast unattested in Japanese, namely a /r/ – /l/ contrast in English. The authors used synthetic stimuli, opposed to natural speech samples, on the assumption that prototypical, idealised synthetic stimuli would provide learners with crucial acoustic information needed to distinguish between /r/ – /l/ tokens. The use of synthesised samples was thought to remove seemingly irrelevant speaker-specific information. Contrary to the traditional approach predictions, the authors were forced to conclude that subjects were largely unsuccessful at learning a non-native /r/ – /l/ contrast from synthetic stimuli and thus directly corroborating the claim that speaker's indexical properties are not only actively processed in speech recognition but that they are also necessary for correct formation of new phonetic categories.

Another body of research giving support to a key role of indexical properties in speech perception comes from research on infant perception (Houston 2008 for a discussion on infant perception). Houston and Juszczyk (2000) found that 7.5-month-old infants learnt to recognise isolated words at the same time encoding speaker-specific information about talkers. They were not able to generalise the knowledge of new items to voices of different sex[2]. Variability in speakers seems to be indispensable for formation of native phonetic categories in first language acquisition. In another study (Houston 1999 reported in Houston 2008), infants were trained to recognise isolated words from a set of dissimilar talkers rather than one voice. The results showed that infants were better at recognising learnt words in passages when trained on multiple talkers than only on one talker. The author concluded that a relatively large distribution of talkers allows infants to

[2] As noted by the reviewer, the experiment by Houston and Jusczyk (2000) may be interpreted as showing that infants need many talkers in order to abstract from their indexical properties rather than, as taken to show in our interpretation, to encode and store talker's individual properties. While this line of reasoning is perfectly legitimate, it is weakened by the results of experiments with adult subjects. If, as claimed by the reviewer, infants learn to abstract from individual properties of a talker, it is difficult to understand why they are not able to do it as adults. The experiments by Mullenix and Pisoni (1990), Goldinger et al. (1991), or Nygaard et al. 1994) demonstrated that familiar voices are processed more effectively than unfamiliar ones.

form more robust and generalizable acoustic representations of words with individual speaker's indexical properties constituting a substantial portion of a stored word specification.

3. Concluding remarks

Speech reaching listeners' ears does not contain only sequences of discrete speech sounds. It is permeated by acoustic information indicative of individual speaker's vocal tract architecture, rate of speaking, and idiosyncratic mannerism in terms of both segment articulation and prosody control. As the evidence reviewed above suggests, these properties are not discarded in speech perception. They are actively processed and stored as episodic traces in long-term memory. This fact has not only practical consequences for methodology of language teaching where variability in an array of talkers will be welcomed as providing opportunities to form more complete generalisations for new sound categories. It is also for crucial importance in revising the applicability of synthetic stimuli in speech perception experiments (Rojczyk 2008 for a discussion). Synthesised speech will of necessity be deprived of indexical properties and thus serve as an incomplete experimental stimulus. Natural speech, on the other hand, will provide an experimenter with natural, information-rich, experimental material for subsequent modification and application in perception experiments.

References

Bradlow, A. R., G. M. Torretta, D. B. Pisoni 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication* 20, 255–272.

Bradlow, A. R., D. B. Pisoni 1999. Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *Journal of the Acoustical Society of America* 106, 2074–2085.

Clopper, C. G., D. B. Pisoni 2008. Perception of dialect variation. In D. B. Pisoni, R. E. Remez (eds.) *The Handbook of Speech Perception*, 313–337. Malden: Blackwell Publishing.

Goldinger, S. D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105, 251–279.

Goldinger, S. D., D. B. Pisoni, J. S. Logan 1991. On the locus of talker variability effects in recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17, 152–162.

Halle, M. 1956. Review of *Manual of Phonology* by C. D. Hockett. *Journal of the Acoustical Society of America* 28, 509–510.

Hockett, C. F. 1955. *Manual of Phonology*. Bloomington: Indiana University.

Houston, D. M. 2008. *The role of talker variability in infant word representations*. PhD Dissertation. The Johns Hopkins University.

Houston, D. M. 2008. Speech perception in infants. In D. B. Pisoni, R. E. Remez (eds.) *The Handbook of Speech Perception*, 417–448. Malden: Blackwell Publishing.

Houston, D. M., P. W. Jusczyk 2000. The role of talker-specific information in word segmentation by infants. *Journal of Experiomental Psychology: Human Perception and Performance* 26, 1570–1582.

Johnson, K. 2008. Speaker normalization in speech perception. In D. B. Pisoni, R. E. Remez (eds.) *The Handbook of Speech Perception*, 363–389. Malden: Blackwell Publishing.

Joos, M. A. 1948. Acoustic phonetics. *Language* 24, 1–136.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, M. Studdert-Kennedy 1967. Perception of a speech code. *Psychological Review* 74, 431–461.

Licklider, J. C. R. 1952. On the process of speech perception. *Journal of the Acoustical Society of America* 24, 590–594.

Local, J. 2002. Variable domains and variable relevance: Interpreting phonetic exponents. Paper presented at the ISCA International Tutorial and Research Workshop on Temporal Integration in the Perception of Speech. Aix-en-Provence, France.

Miller, J. L., F. Grosjean, C. Lomanto 1984. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica* 41, 215–225.

Moon, S. J., B. Lindblom 1994. Interaction between duration, context, and speaking style in English stressed vowels. *Journal of the Acoustical Society of America* 96, 40–55.

Mullenix, J. W., D. B. Pisoni, C. S. Martin 1989. Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America* 85, 365–378.

Mullenix, J. W., D. B. Pisoni 1990. Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics* 47, 379–390.

Murphy, G. L. 2002. *The Big Book of Concepts*. Cambridge: MIT Press.

Nygaard, L. C. 2008. Perceptual integration of linguistic and nonlinguistc properties of speech. In D. B. Pisoni, R. E. Remez (eds.) *The Handbook of Speech Perception*, 390–413. Malden: Blackwell Publishing.

Nygaard, L. C., M. S. Sommers, D. B. Pisoni 1994. Speech perception as a talker-contingent process. *Psychological Science* 5, 42–46.

Pierrehumbert, J. B., R. T. Pierrehumbert 1990. On attributing grammars to dynamical systems. *Journal of Phonetics* 18, 465–477.

Peterson, G. E., H. L. Barney 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24, 175–184.

Pisoni, D. B., S. V. Levi 2007. Representations and representational specificity in speech perception and spoken word recognition. In M. G. Gaskell (ed.) *The Oxford Handbook of Psycholinguistics*, 3–18. Oxford: Oxford University Press.

Rojczyk, A. 2008. *Perception of Polish and English obstruents*. PhD Dissertation, University of Silesia.

Strange, W., S. Dittmann 1984. Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception and Psychophysics* 36, 131–145.

Twaddell, W. F. 1952. Phonemes and allophones in speech analysis. *Journal of the Acoustical Society of America* 24, 607–611.

Uchanski, R. M. 2008. Clear speech. In D. B. Pisoni, R. E. Remez (eds.) *The Handbook of Speech Perception*, 207–235. Malden: Blackwell Publishing.

Vaissiére, J. 2008. Perception of intonation. In D. B. Pisoni, R. E. Remez (eds.) *The Handbook of Speech Perception*, 236–263. Malden: Blackwell Publishing.

Winters, S. J., S. V. Levi, D. B. Pisoni 2006. The role of linguistic competence in cross-linguistic speaker identification. Paper presented at the 80th annual meeting of the Linguistic Society of America, Albuquerque.