SPEECH PERCEPTION IN THE NEXT TEN YEARS: TECHNOLOGICAL SOLUTIONS vs. ACQUIRING ACTUAL SPEECH KNOWLEDGE

Louis C.W. Pols

Institute of Phonetic Sciences
University of Amsterdam, The Netherlands

ABSTRACT

Technological developments will continue to have a strong influence on basic and applied research in phonetic sciences. Solutions that improve performance in speech technology systems, so far have contributed little to our knowledge about human speech communication processes. That is why in the future more speech perception research for its own sake will be required, not just with speech(-like stimuli) under controlled laboratory conditions but also with 'real' speech.

1. INTRODUCTION

In her review for this conference, about 'Phonetics in the next ten years', Keating [2] emphasizes that, instead of *predicting* future research activities, one generally extrapolates from present and past situations. Doing so one can safely say that technological developments will continue to have a strong influence on basic and applied research in phonetics. Nowadays it is much easier to analyze many different aspects of speech, with the consequence that *more* can be measured. This 'more' is both in terms of all kinds of speech characteristics, as well as in terms of different languages, speakers, conditions, and styles. However, measuring more is not necessarily knowing more.

In looking ahead it is wise to look backward as well, in order to have a point of reference for judging progress in acquiring phonetic knowledge. Subsequent International Congresses of Phonetic Sciences (ICPhS) are good occasions for that because of their rather long, four-year, span and because of their emphasis on phonetics.

In my short contribution I would like to emphasize the need to improve our knowledge about the process of human speech perception. In the past, speech perception was a research topic in its own merit, presently it is frequently considered background knowledge or a by-product of research for improving automatic speech recognition, spoken language understanding, and synthesis-by-rule.

2. SPEECH PERCEPTION vs. SPEECH RECOGNITION

The speech databases, used to train those speech recognition systems that are based on neural nets or hidden markov models, provide means to acquire a lot of implicit knowledge.

This knowledge is stored in network structures and transition probabilities. However, most of the time there is no systematic and easily-accessible relation between a certain variable, such as speaker, speaking style, speaking rate, phonemic or sentence context, and the parameters of the network. Despite that, the performance of the most advanced of these systems is surprisingly high and their resistance against various sources of variation is steadily improving.

In a way this is unfortunate, since it does not force researchers to go and study these relations in more detail. Instead, speaker variability is tackled by putting a greater variety of speakers in the training data, context variability is solved by introducing triphone models, rate variation and duration variation is handled by self loops, etc.

The human listener is much more adaptive to all this (systematic) variability and finds ways to normalize. Although, for the time being, technical solutions have been found in the speech recognition domain, our knowledge about how exactly the human listener acts, has hardly been improved. The few researchers that still adhere to formalized acoustic-phonetic knowledge for performing automatic speech recognition have been far less successful, again indicating the complexity of the problem.

3. SPEECH PERCEPTION vs. SPEECH SYNTHESIS

Similar developments can be signalled in text-to-speech synthesis-by-rule [5]. The topic of 'many speakers' and 'different voice characteristics' is immediately shifted aside by choosing one voice or few voices only. Local rate changes are generally not used at all. Still, many segmental, supra-segmental, and linguistic features have to be modelled to make synthetic speech somewhat acceptable.

Only recently some progress has been made in modelling vowel duration [8]. So far the Klatt rules for American English, dating back from the 9th ICPhS in Copenhagen 1979 [3] and before, were the best available.

Our limited knowledge about context-specific dynamic formant changes has not been improved a lot in the last ten years, despite the fact that that knowledge is indispensable for rule synthesis in every single language. A common way to avoid this deficiency is to choose larger basic units such as diphones in which the nearest-neighbor transitions are already incorporated. This means another missed chance to improve our knowledge about how humans produce and perceive these transitions.

Easier access to large and multi-lingual, segmented and labeled (phonetic, prosodic, and linguistic), speech databases, such as becoming available in the ESPRIT projects SAM and Polyglot, will hopefully provide enough data to give it another try [1,6]. This holds not just for the segmental domain but, for instance, also for prosody in order to improve intelligibility and naturalness of synthetic speech. A better understanding of how humans produce and perceive conversational speech, for instance with respect to phonetic and linguistic reduction, stress assignment (given vs. new information), and prosodic phrasing, would also contribute to more natural synthetic speech.

4. SPEECH PERCEPTION FOR ITS OWN SAKE

In my opinion, the next ten years require a renewed and growing interest in studying the basic processes of human speech perception. Speech perception of course has many sides, from perceiving simple basic speech signal attributes such as pitch, duration, and vowel quality, via dynamic attributes such as formant transitions, and aspects of normalization, segregation, and trade-off, to word perception and lexical access. Although the acoustic analysis and perception of 'real' speech might become a fashionable research topic in the next decade [2], this should not prevent us from studying speech and speech-like stimuli under controlled laboratory conditions as well. For instance, if we knew better how context-specific formant transitions are produced [10] and what the variable and invariant components are that determine their role in speech perception [4,7,11], then we would have acquired some generally-applicable universal knowledge. This will certainly contribute also to improved speech technology products. However, that progress might not be astounding and should not be the main reason for doing it. Otherwise, applying acquired knowledge is an excellent way to test whether it is already complete and formalized. Just as ICPhS always had a rather strong link with phonology, the link with psycho-acoustics and hearing, as well as with psycho-linguistics was also apparent in the last few congresses. These domains are excellent bases for studying speech perception as well [9].

5. REFERENCES

[1] FOURCIN, A. et al. (Eds.) (1989), Speech Input and Output Assessment. Multilingual Methods and Standards, Ellis Horwood Lim., Chichester.

[2] KEATING, P.A. (1991), "Phonetics in the next ten years", this volume.

[3] KLATT, D.H. (1979), "Synthesis by rule of segmental durations in English sentences", Proc. 9th ICPhS, Copenhagen, vol. II, 290-297.

[4] PERKELL, J.S. & KLATT, D.H. (Eds.) (1986), Invariance and variability in speech processes, L. Erlbaum Ass., Hillsdale, N.J.

[5] POLS, L.C.W. (1990), "Does improved performance of a rule synthesizer also contribute to more phonetic knowledge?", Proc. ESCA Tutorial Day on Speech Synthesis, Autrans, 49-54.

[6] POLS, L.C.W. (1990), "How useful are speech databases for rule synthesis development and assessment?", Proc. ICSLP-90, Kobe, Vol. 2, 1289-1292.

[7] POLS, L.C.W. & SCHOUTEN, M.E.H. (1987), "Perception of tone, band, and formant sweeps", In: Schouten (Ed.), 231-240.

[8] SANTEN, J.P.H. van & OLIVE, J.P. (1990), "The analysis of contextual effects on segmental duration", Computer Speech and Language 4, 359-390.

[9] SCHOUTEN, M.E.H. (Ed.) (1987), The Psychophysics of speech perception, M. Nijhoff Publ., Dordrecht.

[10] SON, R.J.J.H. van & POLS, L.C.W. (1990), "Formant frequencies of Dutch vowels in a text, read at normal and fast rate", J. Acoust. Soc. Am. 88(4), 1683-1693.

[11] WIERINGEN, A. van & POLS, L.C.W. (1990), "Transition rate-dependent processing of one-formant speech-like stimuli", IFA Proc. 14, 1-16.