PSYCHOPHYSICS, SPEECH PERCEPTION, AND AUTOMATIC SPEECH RECOGNITION

Louis C.W. Pols

Institute of Phonetic Sciences
University of Amsterdam, The Netherlands

ABSTRACT - Research in psychophysics is mainly concentrated on the perception of basic characteristics of relatively simple, stationary, isolated, sounds. Speech, however, is a complex, dynamic, acoustic signal, embedded in context, and linguistically meaningful. It is a temptation to try to bridge the gap between psychoacoustics and speech perception in terms of stimuli, methods, and models used, thus contributing to our knowledge for improving automatic speech recognition as well.

INTRODUCTION

Especially the knowledge-based, or expert, systems in automatic speech recognition (ASR) require a substantial amount of detailed facts about speech and language (e.g. Zue and Lamel, 1986). This involves the signal level as well as the word and sentence level, including such variations as caused by speakers, speaking rate, linguistic and acoustic context, emotion, and acoustic environment. For specific applications, part of that knowledge may be achieved by statistical means. In fact it is surprising to see how efficient knowledge can be represented in hidden Markov chains and can be applied for large-vocabulary, single-speaker, isolated-word, speech recognition (Jelinek et al., 1985). However, it is my firm belief that, in the long run, it is more efficient to understand the many sources of variation, rather than introducing limited domains and describing the variation in such a specific domain in a stochastic way. Every new source of variation then requires a completely new training phase. Because of the use of hugh computer power, this approach so far has been rather successful, whereas the "knowledge-driven approach" stays behind with limited success on partial subtasks only (Glass and Zue, 1986; Pitrelli, 1986; Espy-Wilson, 1986). It seems to be the fate of speech scientists that all their speech knowledge, gathered over the last decades, still seems to be insufficient to make a substantial contribution to automatic speech recognition. We do nevertheless see operational systems, but based on the engineering approach, the stochastic approach, or the artificial intelligence approach.

Unfortunately, I believe that it really is true that we have yet insufficient knowledge about the systematics and (ir)relevance of the variations in the speech signal and the way we perceive all that information. Although we do not necessarily have to imitate the human speech perceiver in order to build a good speech recognizer, such knowledge will certainly show directions towards solving that technical problem. Especially since the human listener is such an excellent pattern, and more specifically speech recognizer. Contrary to any existing recognition system so far, the human listener is rather insensitive to background noise, reverberation or competing voices, to changes from one speaker to the other, to rate and stress variation, to style, dialect, or speaking habits, to grammatical structure, or to changing prosody.

HUMAN PERFORMANCE

Let me just give a few examples of the amazing capabilities of the human listener to adapt to changing listening conditions.

Plomp et al. (1986) recently demonstrated that the speech reception threshold (SRT), for short sentences against a background of noise with a speech-like spectrum, is remarkably resistent against a wide range of different slopes of the communication channel (from -7 to +10 dB/oct), even if the slope varies dynamically. This relative insensitivity of the ear to spectral tilt is a strong indication that it might be more appropriate to use spectral derivatives than absolute spectral values in ASR (Pols and Plomp, 1986).

Another example might be the flexibility with which a human listener can interpret temporal aspects in various contexts, like the duration of the silent interval between /s/ and /l/ in order to differentiate between the Dutch words "slijt" and "splijt" (Pols, 1984), synonymous to the English example "slit/split". There does not seem to be a simple systematic behavior in those various contexts, although it is most certainly not sufficient to use one fixed threshold only.

Another astonishing property of the human observer is his capability to "listen through the noise", not just in terms of signal-to-noise ratio but also in terms of interpreting what might be absent or masked during part of the acoustic event. Pols (1982), for instance, showed that at SNR = -3 dB listeners still do an excellent job in identifying digit sequences, while even at SNR = -9 dB performance is far better than chance, whereas small amounts of noise already strongly deteriorate the performance of ASR-systems. Plomp (1981) gave some nice examples of interrupted natural sounds which were heard continuously when noise was introduced in the gaps. Such experiments, concerning this so-called continuity effect, show that the hearing system is able to restore sound patterns when the duration of the masked portion does not exceed a few hundred of msec; this must be a central rather than a peripheral process. The noise paradigm was also effectively used by us in phoneme identification experiments (Pols and Schouten, 1985). No single ASR-system so far has such pattern interpretation capabilities.

Finally I want to mention some recent listening experiments which we performed in order to elucidate the role of dynamic events, such as rapid formant transitions, in speech perception. We will study the perception of dynamic signals by systematically moving from single- and two-tone sweeps, via single- and multiple-band sweeps, to natural speech segments, by using identification, discrimination, and matching tasks (Pols and Schouten, in press). At this moment we can only conclude that there is a strong interaction between stimulus characteristics and task variables, while an easy interpretation from psycho-physical results to speech perception results does not seem to be possible.

HUMAN AND MACHINE MODELS OF SPEECH RECOGNITION

In Table I a global overview is given of various types of signals, methods, models, and units used in research in psychoacoustics, speech perception, and automatic speech recognition. It is interesting to see that especially at the word recognition level one can notice a trend that models for

automatic speech recognition, like HARPY (Lowerre and Reddy, 1980), LAFS (Klatt, 1986), and TRACE (McClelland and Elman,1986; Elman and McClelland, 1986), and models for human word recognition and lexical access, like the word-initial cohort theory (Marslen-Wilson, 1980), are approaching each other (Pols, in press).

PSYCHOACOUSTICS AND HEARING	SPEECH PERCEPTION	AUTOMATIC SPEECH RECOGNITION
SIGNALS		
-simple	-from single features to natural speech	-complex
-stationary		-dynamic
-isolated		-context embedded
METHODS		
-(masked) thresholds	-discrimination	-speech enhancement
-jnd	-identification	-endpoint detection
-pitch, loudness, timbre, binaural perception	-matching	-distance metrics
	-memory task	-non-linear time normalization
	-lexical decision	-speaker adaptation
	-semantic scaling	-segmentation and labeling
	-word gating	-lexical access
	-shadowing	
MODELS		
-peripheral vs. central processing	-motor theory	-template matching
-time vs. place pitch	-logogen	-acoustic-phonetic approach
-non-linearities	-cohort	-stochastic approach
	-trace	-blackboard, modules
		-Boltzman machines
		-expert systems
		-Trace
		-Feature
		-lexical acccess via midclasses
		-LAFS
UNITS		
-feature	-feature	-feature
	-phoneme	-phoneme
	-word initial cohort	-demisyllable
		-syllable
		-word

Table I. A global overview of signals, methods, models, and units used in research in psychoacoustics, speech perception, and automatic speech recognition.

REFERENCES

ELMAN, J.L. & McCLELLAND, J.L. (1986) "Exploiting lawful variability in the speech wave", In: Perkell and Klatt (Eds.), 360-380.

ESPY-WILSON, C.Y. (1986) "A phonetically based semivowel recognition system, Proc. IEEE-ICASSP86, 2775-2778.

GLASS, J.R. & ZUE, V.W. (1986) "Detection and recognition of nasal consonants in American English", Proc. IEEE-ICASSP86, 2767-2770.

JELINEK, F. + IBM Speech Recognition Group (1985) "A real-time, isolated-word, speech recognition system for dictation transcription", Proc. IEEE-ICASSP85, 23.5.1-23.5.4.

KLATT, D.H. (1986) "The problem of variability in speech recognition and in models of speech perception", In: Perkell and Klatt (Eds.), 300-319.

LOWERRE, B. & REDDY, D.R. (1980) "The Harpy speech understanding system", In: W.A. Lea (Ed.), Trends in speech recognition, (Englewood Cliffs, N.J., Prentice-Hall), 340-360.

McCLELLAND, J.L. & ELMAN, J.L. (1986) "The TRACE model of speech perception", Cognitive Psychology 18, 1-86.

MARSLEN-WILSON, W.D. (1980) "Speech understanding as a psychological process", In: J.C. Simon (Ed.), Spoken language generation and understanding, (D. Reidel, Dordrecht), 39-68.

PITRELLI, J.F. (1986) "Recognition of word-final unstressed syllables", Proc. IEEE-ICASSP86, 2771-2774.

PERKELL, J.S. & KLATT, D.H. (1986) (Eds.) Invariance and variability in speech processes, (Lawrence Erlbaum Ass., Publ., Hillsdale, N.J.).

PLOMP, R. (1981) "Perception of sound signals at low signal-to-noise ratios", In: D.G. Getty and J.H. Howard (Eds.), Auditory and visual pattern recognition, (Hillsdale, Erlbaum), 27-35.

PLOMP, R., ANEMA, P.C. & DIJKHUIZEN, J.N. van (1986) "Towards a hearing aid with multichannel automatic gain control", Proc. 12th ICA, Vol. 1, B4-1.

POLS, L.C.W. (1982) "How humans perform on a connected-digits data base", Proc. IEEE-ICASSP82, 867-870.

POLS, L.C.W. (1984) "Phoneme identification in isolated stimuli and in context", IFA Proc. 8, 33-40.

POLS, L.C.W. (in press) "Interaction between human and machine models of speech recognition", Proc. Symp. on Language Technology, Tilburg, 1985.

POLS, L.C.W. & PLOMP, R. (1986) "How to make more efficient use of the fact that the speech signal is dynamic and redundant", Proc. IEEE-ICASSP86, 1963-1967.

POLS, L.C.W. & SCHOUTEN, M.E.H. (1985) "Plosive consonant identification in ambiguous sentences", J. Acoust. Soc. Amer. 78, 33-39.

POLS, L.C.W. & SCHOUTEN, M.E.H. (In press) "Perception of tone, band, and formant sweeps", Proc. Symp. on Psychophysics and Speech Perception, Utrecht, July 1986.

ZUE, V.W. & LAMEL, L.F. (1986) "An expert spectrogram reader: A knowledge-based approach to speech recognition", Proc. IEEE-ICASSP86, 1197-1200.