Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 21 (1997), 1-10

FLEXIBLE, ROBUST, AND EFFICIENT HUMAN SPEECH RECOGNITION[1]

Louis C.W. Pols

Abstract

In describing human performance in sound perception, in word recognition, in speech understanding, and in dialogue handling, we generally test human limits under controlled conditions and try to understand the underlying mechanisms, however, the human system itself has already been built by nature. In speech and language technology we would like to equal, or perhaps even outrank, human performance, but we will then first have to design the system and we will have to develop the modules according to certain specifications. This paper emphasizes the flexibility, robustness, and efficiency of human performance at various levels and tries to indicate lessons to be learned for designing speech and language technology systems.

1. Introduction

Human speech recognition and understanding is extremely robust to masking, reverberation, and all kinds of other signal distortions. However, a human listener can also easily get distracted from the main speech perception task by other interesting stimuli, or can become tired, or otherwise less attentive. This implies that, on the one hand, under certain conditions an automatic speech understanding system might do better than the human, whereas on the other hand speech and language technology definitely can still learn a lot from human performance.

In a recent paper, Lippmann (1997) concludes that even the presently best single systems for specific tasks, varying from 10-word vocabularies to 65,000-word vocabularies, are still one or more orders of magnitude worse than human performance on similar tasks. He suggests that the human-machine performance gap can be reduced by basic research on improving low-level acoustic-phonetic modeling, on improving robustness with noise and channel variability, and on more accurately modeling spontaneous speech. I see no reason to disagree with him at all, however, let me try to indicate in somewhat more detail a number of reasons why this gap might still exist.

Human listeners generally do not rely on one or a few properties of a specific speech signal only, but use various features that can be partly absent ('trading relations'), a speech recognizer generally is not that flexible. Humans can also quickly adapt to new conditions, like a variable speaking rate, telephone quality speech, or somebody having a cold, using pipe speech, or having a heavy accent. This implies

[1] Slightly adapted version of Pols (1997), Proc. IEEE Workshop on 'Automatic Speech Recognition and Understanding, ASRU'97', Santa Barbara, CA.

that our internal references apparently are not fixed, as they are in most recognizers, but are highly adaptive. Because of our built-up knowledge of speech and language, we can also rather well predict what might come next, in this way making communication much more effficient.

Below I will present a number of these speech perception phenomena and I will try to indicate how they might be useful in speech technology. Unfortunately, I will only rarely be able to demonstrate that by careful implementation of such features, speech recognition performance will actually improve. On the one hand, this is related to our lack of having flexible algorithms available that easily permit us to implement new features (e.g., Sitaram & Sreenivas, 1997), on the other hand this also relates to the fact that simply adding one specific feature, not yet will immediately improve overall system performance, because optimization of *all* elements of a system is required in optimal cooperation (Bourlard et al., 1996; Pols et al., 1996).

2. Aspects of human performance

Human performance in speech perception and recognition is surprisingly flexible, robust, and efficient. For an overview, see for instance Pisoni & Luce (1986) or Allen (1994). The sections below are meant to create some more awareness of these phenomena and their underlying mechanisms, whereas at the same time speech technologists might see good opportunities to implement certain elements for improving the performance of their speech technology systems.

2.1. Robustnes to noise and reverberation

The performance of speech recognizers trained in quiet generally starts to degrade substantially already at signal-to-noise ratios (SNR) of +10 dB and less (Lippmann, 1997), whereas human speech intelligibility (or word error rate) then is not yet degraded at all. Also the level of (human) performance of course depends on such aspects as the size of the vocabulary and the native language of the speaker and the listener. At about -10 dB SNR all speech becomes unintelligble even for very limited vocabularies, such as the digits or the spelling alphabet (Steeneken, 1992). For a difficult word vocabulary such as CVC nonsense words the score from unintelligible to 100% correct covers a range of signal-to-noise ratios of about 20 dB, roughly from -9 to +12 dB. At SNR= -3 dB single digits and triplets in English are still correctly understood with less than 1% error (Pols, 1982).

In Pols (1983) we studied consonant intelligibility and confusibility under various conditions of noise (noise with a speech-like spectrum, and (low-pass filtered) pink noise, SNR from +15 to -6 dB) and reverberation (T = 0, 0.5, 1, and 1.5 s). The theoretical and practical relations between the effect of noise and reverberation and speech intelligibility are nicely represented in the speech trasmission index (STI) concept (Houtgast et al., 1980).

2.2. Robustness to spectral and temporal distortions

Most automatic speech recognizers are trained to learn phoneme, triphone, or word templates. Recognition is then based on shortest distance or greatest similarity. Sometimes speaker adaptation is applied. Also human beings learn by being confronted with many examples, however, their templates seem to be much more flexible and adaptable (Grossberg, 1986). High-pitched small-headed youngsters seem to have little difficulty to understand their low-pitched big-headed fathers. Telephone-quality speech (300-3400 Hz) is not a big challenge for most human listeners. Substantial variability in speaking rate does not seem to bother us a lot.

Actually this relative insensitivity to (spectral) distortions is a prerequisite for a certain type of digital hearing aid to be successful. Ordinary hearing aids amplify the input signal, sometimes after applying a (fixed) filter. Unavoidably this implies that the desirable signal (speech) is amplified just as much as the undesirable signal (noise, competing speech), thus not actually improving the signal-to-noise ratio. By dividing the spectrum in a number of frequency bands and then selectively amplifying certain bands (where the SNR is good) and neglecting others (where SNR is poor already) one can get a much better result. However, this implies that the average speech spectrum continuously changes form. It appears that humans are rather insensitive to that. This can be nicely illustrated by an experiment in which the slope of the amplitude-frequency response is slowly modulated. Sinusoidal variations of the slope from -5 to +5 dB/oct, with frequencies from 0.25 up to 2 Hz, had remarkably little influence on the speech reception threshold (SRT) of sentences in noise (van Dijkhuizen et al., 1987).

Such slope modulations still leave the variation in the envelope per frequency band intact. This temporal envelope contains the information that is essential for the identification of phonemes, syllables and words. Disturbances like noise and reverberation reduce the temporal modulation depth. By applying envelope filtering, Drullman et al. (1994) studied the effect of temporal smearing on speech intelligibility. For low-pass cutoff frequencies above 4 Hz the phoneme intelligibility for CVC and VCV test words was hardly degraded. The same is true for high-pass cutoff frequencies lower than 8 Hz. Apparently we are not very sensitive to temporal smearing.

Ter Keurs et al. (1993) performed a similar study on spectral envelope smearing. Only when the spectral energy is smeared over a bandwidth wider than one-third octave, the masked SRT starts to degrade. This indicates that the intelligibility primarily depends on the global shape of the spectral envelope and not so much on the fine detail.

Flanagan (1972) already demonstrated the low human sensitivity for formant bandwidth, the difference limen being 20 to 40% for one-formant vowel-like stimuli. On the other extreme we can mention sinus speech (Remez et al., 1981), in which formants are reduced to pure tones.

2.3. Auditory modelling

Neuro-mechanical signal processing in the peripheral auditory system is so complex that it does not make much sense to try to imitate that process in ASR front-end modelling, apart from its functionality. Why to worry about the non-flat frequency response of the middle ear, limited spectral resolution of the basilar membrane, limited dynamic range and saturation of the haircells, non-linearities like two-tone suppression, combination tones and lateral inhibition, active elements like the Kemp-echo, co-modulation, profile analysis, or low pitch, if bandfilter analysis, PLP, or MFCC seem to perform rather well already? Of course certain aspects might become more relevant if optimal feature extraction is required. It is probable that higher robustness can be achieved by careful selection of the spectro-temporal features, and that prosody-driven recognizers will indeed increase performance (see also sect. 2.8).

2.4. Multiple features

One of the biggest distinctions between machine recognition and human perception, is the flexible multi-feature approach taken by humans versus the fixed and limited feature approach by pattern recognition machines. A frequently quoted example is the study by Lisker (1978), who showed that the voicing distinction between American English 'rapid' and 'rabid' can independently be controlled by a variety of some 15 different acoustic features. Especially the Haskins group has performed many trading-relations and multiple-cue experiments (e.g., Repp, 1982), see also Nearey (1997). During that same period in time also the invariance theory was popular (Blumstein & Stevens, 1980). The book edited by Perkell & Klatt (1986) is a nice reflection of these discussions. It is also educational to read in Zue (1985) how variable the features are that are used by an expert spectrogram reader.

Personally I believe that there are neither single most important cues, nor invariant cues, but that the flexible human recognizer is as efficient as possible and uses the most appropriate cues from whatever cues are available. If this is a correct viewpoint, worth to be copied in ASR, then of course this does not make life easier at all! In sect. 2.10 I will emphasize another level of complexity, namely the context-dependency.

2.5. Scale spacing: from global to detailed

Some time ago there was a nice beer advertisement on Dutch TV, starting with a shot from the whole earth taken from an outer orbit satellite, and gradually zooming in to Europe, Holland, the Dutch North Sea coast, the Scheveningen beach, up to a good-looking young lady drinking a cool glass of beer on a terrace. This might be an appropriate metaphor for speech perception as well. If necessary we zoom in to the smallest detail (Smits, 1995), if we can do without it, we limit attention to global features only.

Again a speech recognizer generally learns features at one level of precision only. The frame rate frequently is of the order of 10 - 25 ms, although this might be far too detailed for most slowly-changing sonorants, whereas for a burst onset much higher temporal precision might be required.

2.6. Adaptation, speaker normalization

Human adaptation to different speakers, speaking styles, speaking rates, etc. is almost momentarily. However, most so-called adaptive speech recognizers need sizable chunks of speech to adapt. In Pallett et al. (1995) it was clearly demonstrated that most CSR systems, if not adapted, do much worse for the faster speakers in a group. Adapting to another condition, be it more background noise, another speaker, or a different speaking style, should not require new training, but just a quick adaptation of all models. The idea of making optimal use of parameter dependence recently got more attention, and using tree-based multiscale dependency models might be a good approach (Kannan & Ostendorf, 1997).

2.7. Predictability

Everyday experience, as well as formalized gating (Grosjean, 1980), shadowing (Marslen-Wilson & Tyler, 1981) and silent center experiments (Strange, 1989), and phoneme restoration tasks (Samuel, 1981) tell us that humans are rather good (better than an n-gram language model) in predicting what might come next in the speech stream, in this way easing recognition substantially. One could perhaps say that the perplexity for human listeners is always much lower than for machines. Furthermore, most recognizers are not very good in left-to-right processing and prefer to parse a whole sentence.

A somewhat related problem is that of out-of-vocabulary words, which unavoidably sets the upper limit of word-error-rate performance of any CSR system, whereas humans have little difficulty to understand, interpret, and remember unknown words or new word compounds.

2.8. Prosody-driven recognition

To the best of my knowledge Nöth et al. (1997) are the first claiming that their Verbmobil speech understanding system actually uses prosody, although not yet for word recognition itself but for disambiguating speech understanding. They give the example of "Dann müssen wir noch einen Termin ausmachen" ("Then we still have to fix a date") versus "Dann müssen wir noch einen Termin ausmachen" ("Then we still have to fix another date").

The acoustic parameters responsible for prosody are generally considered to be fundamental frequency, duration, energy, and spectral slope, as a function of time. Next to the common aspects of intonation (sentence type, sentence accent), there are strong indications that in human communication the prosodic structure is also responsible for marking word boundaries, for phrasing, and for specifying the pragmatic discourse structure (e.g., Cutler & Butterfield, 1991; van Donzel & Koopmans-van Beinum, 1996). Another related feature of conversational speech are its disfluencies, including filled pauses (Siu & Ostendorf, 1997). If word stress could be detected consistently or, even better, the weak-strong syllable sequence, this could greatly enhance word recognition performance.

2.9. Duration modeling

Phoneme duration is one of those signal aspects for which a phonetician believes that this could be modelled better than just by a probability density distribution. Speaking style, speaking rate, word stress, local context, position in the word, and position of the word in the sentence, all contribute to the actually realized phoneme duration. Analyzing a large database undoubtedly shows that, both for vowel (Pols et al., 1996) and consonant duration (van Son & van Santen, 1997). Unfortunately, the benefit in terms of increased ASR performance of using that knowledge cannot so easily be demonstrated (Wang, 1997).

2.10. Coarticulation and reduction

In our institute we have paid much attention to the dynamic spectro-temporal events (formant transitions) in speech.

This led, for instance, to a better understanding of the human sensitivity to *vocalic transitions* (van Wieringen & Pols, 1995). Whereas the difference limen (DL) in endpoint frequency for 40-ms tone glides is as low as 30 Hz, it is more than 200 Hz for VC-like stimuli with a short (20 ms) formant transition. This may be another indication that high spectral resolution is not always required and that unique spectral targets are quite useless.

By comparing formant transitions in *normal* and *fast* rate speech for comparable CVC-segments, Pols & van Son (1993) could show that, at least for this male speaker, so-called formant undershoot in the shorter segments of fast-rate speech, did barely happen. Apparently this speaker could easily adapt his speaking style (articulation speed) in such a way that still the vowel target, appropriate for that context, could be reached. Of course, contextual and prosodic conditions caused a lot of (rather systematic) variation in the vowel midpoint formant position reached, but higher speaking rate and shorter duration did add very little to that variability. On the other hand, changing the *speaking style* from *read* to *spontaneous* speech, did cause vowel reduction, more specifically a centralization of mainly F_1 (van Son & Pols, 1996). We have the impression that it might be useful to implement such rather systematic phenomena as specific knowledge in ASR, rather than as variability in training data.

In a similar way, van Bergem (1995) greatly enhanced our insight on *vowel reduction*. He showed that that mechanism is much more a process of contextual assimilation than of centralization. This may have implications for the phone and word models used in ASR. Similarly, the (Dutch) schwa appears to be a vowel without an articulatory target, completely assimilated with its (consonantal) environment. This coarticulatory effect of C_1, C_2, and V on the schwa in ´VC_1/ə/C_2 and C_1/ə/´C_2V nonsense words could very well be modelled. Triphone models are probably rather good in modelling this contextual asimilation, however, they do not distinguish in levels of reduction.

Recently van Son & Pols (1997) drew attention for *consonant reduction* as well. Reduction in consonant identification errors for VCV syllables extracted from spontaneous vs. read speech (for both stressed and unstressed syllables) was compared with the differences in five acoustical measures: segmental duration, spectral center of gravity, intervocalic sound energy difference, intervocalic F_2 slope difference, and the amount of vowel reduction in the syllable kernel. All these acoustic measures appear to be indicators of both vowel and consonant reduction and are all correlated to changes in speaking style and syllable stress. Only for segmental duration and spectral center of gravity we could so far show (in a statistically significant way) that a 'reduction' in these values also correlated to 'reduced' identification. See table 1 for the mean consonant identification results.

	stressed	*unstressed*	*total*
read	14.4	18.0	16.6
spontaneous	22.2	30.5	27.3

Table 1: Mean (22 Dutch listeners) consonant error rate (in percentages) for VCV stimuli (2 x 791 VCV segments extracted from read and spontaneous speech, partly stressed (308), partly unstressed (483)), separated out for speaking style and syllable stress.

2.11. Pronunciation variation

In May 1998 an ESCA workshop on "Modeling pronunciation variation for automatic speech recognition" will be organized in Holland (http://lands.let.kun.nl/ pron-var/). Taking one standard pronunciation from a word lexicon, irrespective of the speaker and the context in which the words occur, is a huge oversimplification. Human lexical search certainly does not work like that. We know about reduction, word boundary effects like deletion and stress clash, allophonic variation, etc. Applying phonotactic rules (Giachin et al., 1991), or extracting detailed pronunciation from large database statistics (Riley & Lolje, 1996; Wang & Pols, 1997), or using separate stressed and unstressed phoneme models (van Kuijk et al., 1996), have so far only let to limited success.

2.12. Word perception models

Psycholinguistics and related domains have provided us with a great variety of speech perception and word recognition models, such as the motor theory (Liberman & Mattingly, 1985), analysis-by-synthesis (Stevens, 1960), quantal theory (Stevens, 1989), logogen model (Morton, 1969), cohort model (Marslen-Wilson & Welsh, 1978), lexical access from spectra (LAFS) (Klatt, 1979), first order context-sensitive coding (ERIS) (Marcus, 1981), autonomous search (Forster, 1976), dual coding (Foss & Blank, 1980), interactive activation TRACE model (McClelland & Elman, 1986), shortlist (Norris, 1994), adaptive learning (Grossberg, 1986), etc. Rather than advocating one best approach, it might be wiser to indicate that HMM-based word recognition is probably not a bad choice after all. It is a kind of analysis-by-synthesis model and it allows for extended unmoderated learning.

What I would like to see added in CSR, is the implementation of more *specific knowledge* that relatively easily could be derived from the speech stream (such as environmental characteristics, speaking style characteristics related to the speaker and the local speaking rate, as well as word characteristics related to word stress, reduction and coarticulation) and that might permit quick adaptation of the model parameters.

As already indicated by G. Doddington at the ARPA Spoken Language System Technology Workshop in 1995 in Austin, TX, including speech understanding in the word recognition process beyond a simple n-gram model, would also enhance CSR performance.

2.13. Other modalities

Speech is definitely an acoustic signal, however, speech communication is not necessarily limited to the auditory mode only, unless the communication channel forces one to do so, like in telephone speech. Arm and body gestures, facial expressions, eye blinks, all add to the communicative situation and may influence the interpretation of what was said. Audio-visual synthesis ("talking faces") is getting more and more popular, if sign language symbols have to be transmitted, the visual modality is of course also unavoidable, but also bimodal ASR is starting to get some attention. At the ESCA Workshop on "Audio-visual speech processing AVSP'97" (Benoit & Campbell, 1997) there was for instance a session on "Automatic recognition of audio-visual speech".

3. Conclusions

This potpouri of observations concerning the flexibility, robustness, and efficiency of human speech perception and word recognition, unfortunately cannot be a manual for ASR-best-practice at all. Much of the apparent systematicity in human perception, either cannot be implemented in present-day recognizers at all, or, if implementable, generally does not lead to any improved performance. So, should speech scientists and speech technologists simply stop trying to understand each other and to learn from each other? Of course not, we should join forces, have more sessions like the one on 'Lessons learned from human speech recognition system' in which this paper was presented at the IEEE Workshop on Speech Recognition and Understanding ASRU'97, Santa Barbara, CA (Furui et al., 1997), and evaluate and compare analytically human and system behavior.

4. References

Allen, J.B. (1994), "How do humans process and recognize speech?", *IEEE Trans. Speech Audio Proc.* **2(4)**, 567-577.

Benoit, C. & Campbell, R. (Eds.) (1997), *Proc. of the ESCA Workshop on Audio-visual speech processing. Cognitive and computational approaches, AVSP'97*, Rhodes.

Bergem, D. van (1995), *Acoustic and lexical vowel reduction*, Ph.D thesis, Univ. of Amsterdam, Studies in Language and Language Use 16.

Blumstein, S.E. & Stevens, K.N. (1980), "Perceptual invariance and onset spectra for stop consonants in various vowel environments", *J. Acoust. Soc. Am.* **67**, 648-662.

Bourlard, H., Hermansky, H. & Morgan, N. (1996), "Towards increasing speech recognition error rates", *Speech Communication* **18**, 205-231.

Cutler, A. & Butterfield, S. (1991), "Durational cues to word boundaries in clear speech: A supplementary report", *Speech Communication* **10**, 335-353.

Dijkhuizen, J.N. van, Anema, P.C. & Plomp, R. (1987), "The effect of varying the slope of the amplitude-frequency response on the masked speech-reception threshold of sentences", *J. Acoust. Soc. Am.* **81(2)**, 465-469.

Donzel, M.E. van & Koopmans-van Beinum, F.J. (1996), "Pausing strategies in discourse in Dutch", *Proc. ICSLP'96*, Philadelphia, Vol. 2, 1029-1032.

Drullman, R., Festen, J.M. & Plomp, R. (1994), "Effect of reducing slow temporal modulations on speech perception", *J. Acoust. Soc. Am.* **95**, 2670-2680.

Flanagan, J.L. (1972), *Speech analysis synthesis and perception*, Springer Verlag, Berlin, 2nd edition.

Forster, K.I. (1976), "Accessing the mental lexicon", In: R.J. Wales & E. Walker (Eds.), *New approaches to language mechanisms*, North-Holland, Amsterdam, 257-287.

Foss, D.J. & Blank, M.A. (1980), "Identifying the speech codes", *Cognitive Psychology* **12**, 1-31.

Furui, S., Juang, B.-H. & Chou, W. (Eds.) (1997), *Proc. of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'97*, IEEE Signal Processing Society, Piscataway, NJ.

Giachin, E.P., Rosenberg, A.E. & Lee, C.-H. (1991), "Word juncture modeling using phonological rules for HMM-based continuous speech recognition", *Computer Speech and Language* **5**, 155-168.

Grosjean, F. (1980), "Spoken word recognition processes and the gating paradigm", *Perception & Psychophysics* **28(4)**, 267-283.

Grossberg, S. (1986), "The adaptive self-organization of serial order in behavior: Speech, language, and motor control", In: E.C. Schwab & H.C. Nusbaum (Eds.), *Pattern recognition by humans and machines*, Vol. I, Speech perception, Academic Press, Inc., Orlando, Chapter 6, 187-294.

Houtgast, T., Steeneken, H.J.M. & Plomp, R. (1980), "Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics", *Acustica* **46**, 60-72.

Kannan, A. & Ostendorf, M. (1997), "Modeling dependency in adaptation of acoustic models using multiscale tree processes", *Proc. Eurospeech'97*, Rhodes, Vol. 4, 1863-1866.

Keurs, M. ter, Festen, J.M. & Plomp, R. (1993), "Effect of spectral envelope smearing on speech perception. II", *J. Acoust. Soc. Am.* **93**, 1547-1552.

Klatt, D.H. (1979), "Speech perception: A model of acoustic-phonetic analysis and lexical access", *Journal of Phonetics* **7**, 279-312.

Kuijk, D. van, Heuvel, H. van den & Boves, L. (1996), Using lexical stress in continuous speech recoginition for Dutch", *Proc. ICSLP'96*, Philadelphia, Vol. 3, 1736-1739.

Liberman, A.M. & Mattingly, I.G. (1985), "The motor theory of speech perception revised", *Cognition* **21(1)**, 1-36.

Lippmann, R.P. (1997), "Speech recognition by machines and humans", *Speech Communication* **22**, 1-15.

Lisker, L. (1978), "Rapid vs rabid: A catalogue of acoustic features that may cue the distinction", *Haskins Labs, Status Report on Speech Research* **SR-54**, 127-132.

Marcus, S.M. (1981), "ERIS - context sensitive coding in speech perception", *Journal of Phonetics* **9**, 197-220.

Marslen-Wilson, W.D. & Tyler, L.K. (1981), "Central processes in speech understanding", *Philosophical Transactions of the Royal Society of London*, **B 295**, 317-332.

Marslen-Wilson, W.D. & Welsh, A. (1978), "Processing interactions and lexical access during word recognition in continuous speech", *Cognitive Psychology* **10**, 29-63.

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

Morton, J. (1969), "Interaction of information in word recognition", *Psychological Review* **76**, 165-178.

Nearey, T.M. (1997), "Speech perception as pattern recognition", *J. Acoust. Soc. Am.* **101(6)**, 3241-3254.

Norris, D. (1994), "SHORTLIST: A connectionist model of continuous speech recognition", *Cognition* **52**, 189-234.

Nöth, E., Batliner, A., Kießling, A., Kompe, R. & Niemann, H. (1997), "Suprasegmental modelling", *Informal Proc. NATO ASI on Computational models of speech pattern processing*, St. Helier, Jersey Channel Islands.

Pallett, D.S., Fiscus, J.G., Fisher, W.M., Garofolo, J.S., Lund, B.S., Martin, A. & Przybocki, M.A. (1995), "1994 Benchmark tests for the ARPA Spoken Language Program", *Proc. ARPA Spoken Language System Technology Workshop*, Austin, TX, 5-36.

Perkell, J.S. & Klatt, D.H. (Eds.) (1986), *Invariance and variability in speech processes*, Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ.

Pisoni, D.B. & Luce, P.A. (1986), "Speech perception: Research, theory, and the principal issues", In: E.C. Schwab & H.C. Nusbaum (Eds.), *Pattern Recognition by humans and machines*, Vol. I, Speech perception, Academic Press., Inc., Orlando, Chapter 1, 1-50.

Pols, L.C.W. (1982), "How humans perform on a connected-digits data base", *Proc. IEEE-ICASSP'82*, Paris, Vol. 2, 867-870.

Pols, L.C.W. (1983), "Three-mode principal component analysis of confusion matrices, based on the identification of Dutch consonants, under various conditions of noise and reverberation", *Speech Communication* **2(4)**, 275-293.

Pols, L.C.W. (1986), "Analysis and perception of dynamic events and of reduction phenomena in speech", In: W. Ainsworth & S. Greenberg (Eds.), *Proc. ESCA Workshop on the Auditory basis of speech perception*, Keele, UK, 17-22.

Pols, L.C.W. (1997), "Flexible human speech recognition", *Proc. of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'97*, IEEE Signal Processing Society, Piscataway, NJ, 273-283.

Pols, L.C.W. & Son, R.J.J.H. van (1993), "Acoustics and perception of dynamic vowel segments", *Speech Communication* **13**, 135-147.

Pols, L.C.W., Wang, X. & Bosch, L.F.M. ten (1996), "Modelling of phone duration (using the TIMIT database) and its potential benefit for ASR", *Speech Communication* **19(2)**, 161-176.

Remez, R.E., Rubin, P.E., Pisoni, D.B. & Carrell, T.D. (1981), "Speech perception without traditional speech cues", *Science* **212**, 947-950.

Repp, B. (1982), "Phonetic trading relations and contexts effects: New experimental evidence for a speech mode of perception", *Psychological Bulletin* **92**, 81-110.

Riley, M.D. & Ljolje, A. (1996), "Automatic generation of detailed pronunciation lexicons", In: C.-H. Lee, F.K. Soong & K.K. Paliwal (Eds.), *Automatic speech and speaker recognition. Advanced topics*, Kluwer Academic Publishers, Boston, 285-301.

Samuel, A.G. (1981), "Phonemic restoration: Insights from a new methodology", *Journal of Experimental Psychology: General* **110**, 474-494.

Sitaram, R.N.V. & Sreenivas, T. (1997), "Incorporating phonetic properties in hidden Markov models for speech recognition", *J. Acoust. Soc. Am.* **102(2)**, 1149-1158.

Siu, M. & Ostendorf, M. (1997), "Variable n-gram language modeling and extensions for conversational speech", *Proc. Eurospeech'97*, Rhodes, Vol. 5, 2739-2742.

Smits, R.L.H.M. (1995), *Detailed versus gross spectro-temporal cues for the perception of stop consonants*, Ph.D thesis, Technological Univ. Eindhoven.

Son, R.J.J.H. van & Pols, L.C.W. (1996), "An acoustic profile of consonant reduction", *Proc. ICSLP'96*, Philadelphia, Vol. 3, 1529-1532.

Son, R.J.J.H. van & Pols, L.C.W. (1997), "The correlation between consonant identification and the amount of acoustic consonant reduction", *Proc. Eurospeech'97*, Rhodes, Vol. 4, 2135-2138.

Son, R.J.H. van & Santen, J.P.H. van (1997), "Strong interaction between factors influencing consonant duration", *Proc. Eurospeech'97*, Rhodes, Vol. 1, 319-322.

Steeneken, H.J.M. (1992), *On measuring and predicting speech intelligibility*, Ph.D. thesis, Univ. of Amsterdam.

Stevens, K.N. (1960), "Toward a model for speech recognition", *J. Acoust. Soc. Am.* **32**, 47-55.

Stevens, K.N. (1989), "On the quantal nature of speech", *Journal of Phonetics* **17**, 3-45.

Strange, W. (1989), "Dynamic specification of coarticulated vowels spoken in sentence context", *J. Acoust. Soc. Am.* **85**, 2135-2153.

Wang, X. (1997), *Incorporating knowledge on segmental duration in HMM-based continuous speech recognition*, Studies in Language and Language Use 29, Ph.D thesis, Univ. of Amsterdam.

Wang, X. & Pols, L.C.W. (1997), "Word juncture modelling based on the TIMIT database", *Proc. Eurospeech'97*, Rhodes, Vol. 5, 2407-2410.

Wieringen, A. van & Pols, L.C.W. (1995), "Discrimination of single and complex consonant-vowel- and vowel-consonant-like formant transitions", *J. Acoust. Soc. Am.* **98(3)**, 1304-1312.

Zue, V.W. (1985), "The use of speech knowledge in automatic speech recognition", *Proc. IEEE* **73(11)**, 1602-1615.