FLEXIBLE, ROBUST, AND EFFICIENT HUMAN SPEECH PROCESSING VERSUS PRESENT-DAY SPEECH TECHNOLOGY

Louis C.W. Pols
Institute of Phonetic Sciences / IFOTT, University of Amsterdam

ABSTRACT

Present-day speech technology systems try to perform equally well or preferably even better than humans under specific conditions. For more complex tasks machines frequently show degraded performance, because their flexibility, robustness and efficiency is lower than that of humans. In order to better understand the system limitations and perhaps further improve system performance, one can try to learn from human behavior and imitate its functionality, without plain duplication. This paper discusses a number of characteristics of human speech processing and compares these with system performance. It is argued that phonetic sciences and speech technology can mutually benefit from each other if they use similar data and similar representations. R. Moore [25] used for this approach the appropriate term *Computational Phonetics*.

1. INTRODUCTION

Whenever a discussion starts about implementing specific (phonetic or linguistic) knowledge in (speech) technological applications, always the metaphor about birds versus airplanes pops up. Planes don't flap wings, so why should speech recognizers have ears [15]? In a way this is also the theme of my keynote address: Are humans indeed much better than machines in processing speech and what can we learn from them to improve the performance of speech technology systems? And more specifically, given the present conference, can basic research in phonetics help speech technology?

Indeed I believe that humans are much better speech communicators than machines, they are far more flexible, robust, and efficient. However, humans are also lazy, get tired or bored, can be pre-occupied, have strong expectations, generally only know one language, etc.. For all these and other reasons present-day speech-technology systems can, under certain conditions, do better than humans. Think for instance about a 24-hours speaker-independent telephone or credit card number recognizer operating over any telephone line, or a never tired or irritated flight or subway announcer using canned speech plus some rule synthesis.

Another hot discussion item concerns the concept of 'knowledge'. Is good old-fashioned phonetic or phonological knowledge expressed in regular expressions, superior or inferior to probability distributions derived from an annotated speech database? Of course it all depends on the validity of the data and upon their usefulness for certain applications. In a formant-based rule synthesizer regular expressions might be very helpful, whereas in an HMM-based speech recognizer probabilistic knowledge might be much more easily implementable.

Below I will present, in a number of sections, various aspects of human speech processing. I will indicate its capabilities and limitations, and I will try to point out how this knowledge might be used to help to improve speech technology.

2. HOW GOOD IS HUMAN AND MACHINE SPEECH RECOGNITION?

Undoubtedly, the performance of speech recognition, understanding, and dialogue systems has greatly improved since the early days of DTW-based isolated-word recognizers. DARPA and NIST officials [7, 29, 30] are very good in showing us how impressive progress has been over the years for ever-more difficult tasks, from the TI-digits and the spelling alphabet, via the 1,000-word naval resource management (RM) database, the air travel information (ATIS) database, the read aloud Wall Street Journal (WSJ), later extended to many more newspapers in the North American Business (NAB) news, and now moving towards truely conversational speech in TVshows (Broadcast News, including the Marketplace broadcast) and over the telephone (Switchboard and Callhome, also in other languages than English). Also in Europe mono- and multi-lingual speech databases become more and more common for training and testing, such as Eurom, Polyphone, Speechdat Car, SALA (SpeechDat across Latin America), Albayzin, BDLEX, Babel, Verbmobil, read aloud Le Monde, travel information calls, and the Corpus of Spoken Dutch [28].

Lippmann [24] has provided an interesting comparison between human and machine performance in terms of word error rate for 6 different tasks, from TI connected digits to phrases from Switchboard telephone conversations, all in the talker-independent mode. Table 1 gives an overview of the best results that he quotes:

corpus	*description*	*vocabul. size*	*recogn. perplex.*	*% word error*	
				machine	*human*
TI digits	read digits	10	10	0.72	0.009
Alphabet	read letters	26	26	5	1.6
Resource Management	read sentences	1,000	60 - 1,000	17	2
NAB	read sentences	5,000 - unlimited	45 -160	6.6	0.4
Switchboard CSR	spontaneous telephone conversations	2,000 - unlimited	80 -150	43	4
Switchboard wordspotting	idem	20 keywords	-	31.1	7.4

Table 1. Summary of the word (or digit string) error rates for humans and for the best performing machines, from [24].

He concludes that even the presently best single systems for specific tasks, varying from 10-word 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.

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 to somebody having a cold, using pipe speech, or having a heavy accent. This implies that our internal references apparently are not fixed, as they are in most recognizers, but are highly adaptive. Because of our built-in knowledge of speech and language we can also rather well predict what might come next, in this way making communication much more efficient [33].

In sect. 4.2 we will discuss another aspect of the difference between human and machine performance, namely the impressive human robustness to noise, level variation, spectral distortion, reverberation, rate change, variable styles and emotions, etc..

3. HOW INTELLIGIBLE IS MACHINE-GENERATED SPEECH?

Machine-generated speech can be produced in many different ways and for many different applications. Using concatenative canned speech at word level produces highly intelligible and almost natural sounding utterances for small-vocabulary applications, such as announcement systems or a speaking clock. Unlimited vocabulary speech synthesis is possible through formant synthesis by rule, but its quality and intelligibility is far from perfect. The better the rules, the higher the quality will be. One compromis is the use of diphones, either (LPC-) parameterized, or using the original waveform plus PSOLA for pitch and duration manipulations. Concatenative units of variable size, taken upon demand from a large speech corpus, are the latest fashion and can produce good quality speech [5]. Still, for each speaking style and for each new speaker type, another corpus is required, unless the source and filter characteristics of these concatenative units can be modified at will [20, 34]. This is still a serious research area.

On the one hand speech synthesizers can already be used to help visually handicapped people to read aloud the newspaper for them [4], on the other hand the intelligible pronunciation of any name, address, and telephone number is still a challenging task.

Last November 1998 an interesting synthesis evaluation took place during the ESCA Workshop on Speech Synthesis in Jenolan Caves, Australia. Some 60 systems, from 39 different providers and research labs, covering 18 different languages, were offered for evaluation by the over 100 workshop participants themselves.

Mainly because at least 3 systems per language were required for a proper comparison, ultimately 42 systems in 8 languages actually participated in the test (see Table 2). The (preferably native) subjects listened to all available systems in that specific language, while these synthesizers produced up to 3 different types of text: newspaper sentences, semantically unpredictable sentences, and telephone directory entries. Software was developed and recordings were made (under controlled conditions: in a short period of time, previously unknown texts had to be generated) that allowed running this listening experiment on site on some 12 different PCs. Too many things went somewhat wrong in this first large scale test in order to allow to make any serious comparison, the workshop participants furthermore agreed not to make any individual results public, but it was perfectly clear that more of such tests are required and that no system was perfect yet! To underscore this statement, let me just mention that most systems had an 80% or less score on semantically unpredictable sentences. These are short and rather simple sentences of the type 'Het oog kent het paard dat blijft' (The eye knows the horse that stays) [1]. They consist of high-frequent words only and should not be a real challenge to present-day synthesizers anymore. Still, this first large-scale synthesis evaluation was most valuable and should get follow-ups.

system	*speakers*			*texts*		
	male	*female*	*both*	*news*	*SUS*	*teldir*
American English	8	-		8	6	6
American English	-	5		5	5	3
British English	4	-		4	3	na
German	7	-		7	6	4
German	-	3		3	3	2
French			3	3	na	na
Dutch			2	2	2	2
Spanish			3	3	na	2
Chinese			3	3	na	na
Japanese			4	4	na	3

Table 2. Some information about the 42 systems that actually were evaluated at the workshop. This concerns: language; number of male and/or female speakers; test material: newspaper sentences, semantically unpredictable sentences, or telephone directory entries; 'na' indicates 'not available' for that language. Systems on one row were compared against each other.

4. WHAT KIND OF EARS DO RECOGNIZERS NEED?

I consider the pre-processor that transforms the speech input signal into a parameter vector for further processing and recognition, to be the recognizer's ear. Any proper preprocessing done at this level that will improve the recognizer's performance or its robustness will be advantageous. I consider this to be true, even if corrections at other levels, such as a hybrid approach, or language modeling, could achieve similar performance.

4.1 Sensitivity for stationary and dynamic signals

The human peripheral and central hearing system has a number of characteristics that are worth to be taken into account [35]. The ear has a certain sensitivity for stationary and dynamic audio signals expressed in terms of detection thresholds and just noticeable differences.

For instance the difference limen for formant frequency is 3-5% for a *stationary* synthetic one-formant stimulus; for formant bandwidth this is only 20-40%. Pitch discrimination under such experimental conditions is rather good (better than 0.5%), but quickly degrades under more realistic conditions. For more details see Table 3.

Perceptual data for *dynamic* and thus more speech-like signals are rare. Van Wieringen & Pols [55] showed that the difference limen for an initial formant transition is as high as 230 Hz for short (20-ms) transitions, but becomes better the longer the transition and the less (spectrally) complex the signal, see Fig. 1.

phenomenon	*threshold/jnd*	*remarks*
threshold of hearing	0 dB at 1000 Hz	frequency dependent
threshold of duration	constant energy at 10 - 300 ms	Energy = Power x Duration
frequency discrimination	1.5 Hz at 1000 Hz	more when < 200 ms
intensity discrimination	0.5 - 1 dB	up to 80 dB SL
temporal discrimination	[a] 5 ms at 50 ms	duration dependent
masking	psychophysical tuning curve	
pitch of complex tones	low pitch	many peculiarities
gap detection	[a] 3 ms for wide-band noise	more at low freq. for narrow-band noise
formant frequency	3 - 5 %	one formant only < 3 % with more experienced subjects
formant amplitude	[a] 3 dB	F2 in synthetic vowel
overall intensity	[a] 1.5 dB	synthetic vowel, mainly F1
formant bandwidth	20 - 40 %	one-formant vowel
F0 (pitch)	0.3 - 0.5 %	synthetic vowel

Table 3. Detection thresholds and jnd (just noticeable differences) for stationary signals and multi-harmonic, single-formant-like periodic signals.

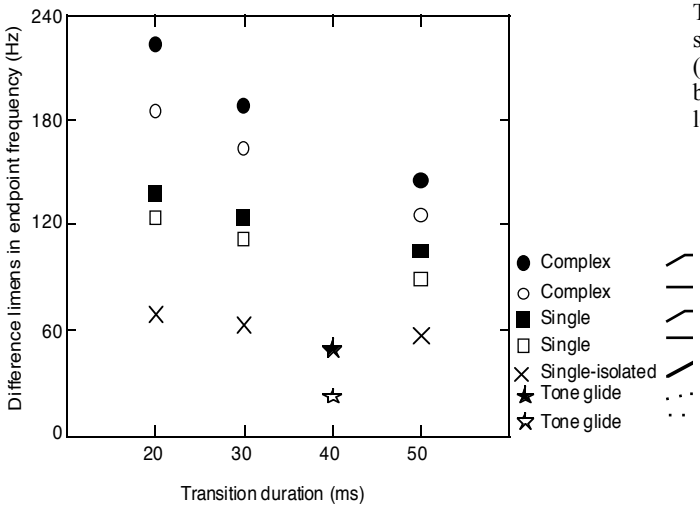

Figure 1. Difference limen, at variable transition duration, in onset or offset frequency (Hz), for initial or final transitions, respectively, of tone sweeps, and of single or complex transitions, in isolation or with a steady state. DL data are averaged over 4 subjects [54].

4.2 Robustness to degraded speech

One way to look at speech is in terms of a time-modulated signal in a number of frequency bands. This envelope modulation exemplifies the existence of words, syllables and phonemes. The modulation spectrum of speech shows a maximum at around 4 Hz. We can degrade speech by temporally smearing out these modulations [10]. Human listeners appear not to be very sensitive to such temporal smearing.

Speech segments do have a power spectrum of which the envelope can also be seen as a modulated signal. Also this spectral envelope can be smeared out, thus reducing spectral contrasts, by using wider and wider filters, say from 1/8 to 2 octaves. Only when the spectral energy is smeared over a bandwidth wider than 1/3 octave, the masked Speech Reception Threshold, a measure for speech intelligibility, starts to degrade [22].

The human ear is also remarkably insensitive (or easily adaptable?) to another type of spectral distortion in which the average speech spectrum continuously changes form. Sinusoidal variations of the spectral slope of the speech signal from -5 to +5 dB/oct, with frequencies from 0.25 to 2 Hz, have remarkably little influence on the SRT of sentences in noise [11]. This insensitivity is actually a requirement for a certain type of digital hearing aid to be successful since these systems continuously amplify frequency bands with a favorable SNR and attenuate the other frequency bands. This implies that the average speech spectrum continuously changes form. It appears that humans are rather insensitive to that. I am afraid that on the other hand speech recognizers are extremely sensitive to such transformations!

I am not argueing here that, because of the above results, recognition preprocessors should use less spectral or temporal resolution, or should use only differential measures. However, I do argue for more flexibility in pre-processing and in feature extraction.

4.3 Robustness 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 [24], 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

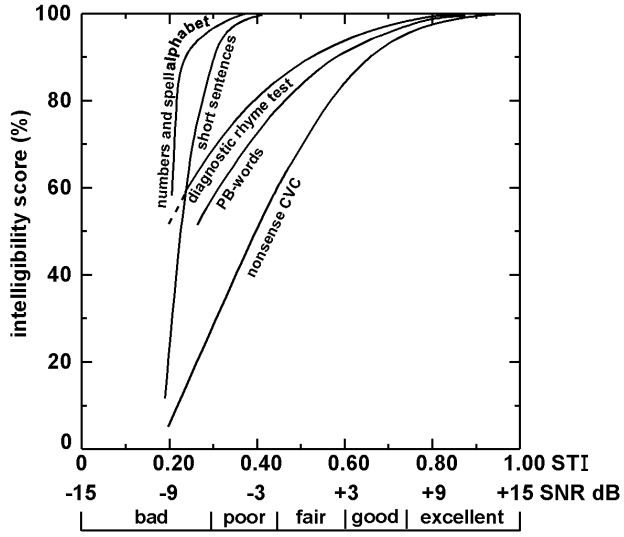

Figure 2. Intelligibility of various word types as a function of the signal-to-noise ratio (SNR) for noise with a speech-like spectrum. The Speech Transmission Index (STI) is also indicated [50].

as the size of the vocabulary and the native language of the speaker and the listener.

At about -10 dB SNR all speech becomes unintelligible even for very limited vocabularies, such as the digits or the spelling alphabet [50]. 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 (see Fig. 2). At SNR = -3 dB single digits and triplets in English are still correctly understood with less than 1% error [32].

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/or reverberation (T = 0, 0.5, 1, and 1.5 s) [31]. Also under such conditions consonant intelligibility starts to degrade at SNR **£** +10 dB. The theoretical and practical relations between the effect of noise and reverberation and speech intelligibility are nicely represented in the speech transmission index (STI) concept based on the Modulation Transfer Function [18].

4.4 Filter characteristics

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 modeling, 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 sect. 5.7.

Hermansky [15] has been especially productive in suggesting and testing various spectro-temporal analysis procedures, such as PLP, RASTA, the use of multi-bands for noisy speech, and most recently TRAPS [16].

5. WHAT KIND OF (PHONETIC) KNOWLEDGE COULD RECOGNIZERS TAKE INTO ACCOUNT?

It is a lost battle to try to return to the old days of knowledge-based recognition (e.g., [56]), however, this should not prevent us from considering specific phonetic and linguistic knowledge that might be implementable in probabilistic recognition and thus hopefully will improve performance. As the title of my presentation indicates, human recognition is flexible, robust and efficient, and it would not hurt recognition machines to have more of these characteristics as well.

It always strikes me that many rather consistent speech characteristics are most of the time totally neglected in speech recognition. Let me mention a few:

- pitch information
- durational variability
- spectral reduction and coarticulation
- quick adaptation to speaker, style and communication channel
- communicative expectation
- multi-modality
- binaural hearing

If you permit me to give a caricature of present-day recognizers, then these machines are trained with all the speech, speaker, textual and environmental variability that may occur under the application in mind, thus giving the system a lot of global knowledge without understanding all the inter-relations. Furthermore, the input is monaural and unimodal and the pitch extractor does not work. Subsequently the recognizer performs rather well on average behavior and does poorly on any type of outlier, be it an unknown word, or a non-native speaker, or a fast speaker, or one with a cold, or a whispered input. The system does not know, or at least is not yet able to use that knowledge, that most question phrases have a rising pitch contour, that in fast speech almost all segments are shorter, that new information is stressed, that actual pronunciation deviates in predictable ways from the normative form given in the lexicon, etc..

It is certainly worth trying to study whether certain local characteristics could be assigned to incoming speech, in order to fine-tune the recognition system and thus hopefully improve its performance.

Such local characteristics should preferably be derivable from the speech signal as such, without knowing yet the word sequence. So, this could be the sex of the speaker, the local speaking rate, the clearness or nasality of articulation, the local prominence, etc.. Once something like an N-best recognition is achieved, another level of post-processing is possible, based on the given word sequence and the potential meaning. At this level one can think of phrase-final lengthening and other boundary markers, poly-syllabic shortening, consonant cluster compression, r-coloring, assimilation, coarticulation, and reduction up to complete deletion, accentual lengthening, lexical stress, accent-lending pitch movements, consequences of stress clash and other rhythmic phenomena, dialectical and speaking style adaptation, etc..

It is of course true that in phone or triphone models a number of the above phenomena are at least partly covered in a probabilistic way, but any consistent behavior is not.

5.1 Durational variability

In order to make this whole discussion slightly more specific, let me present some data on segmental durational variability. One of my former Ph.D. students, Xue Wang, carefully studied the durational phone characteristics from all training sentences in the TIMIT database [57] and incorporated that knowledge in the post-processing rescoring phase of an HMM-based recognizer [37,53]. We studied 11 attributes, but finally choose, for practical reasons, 4 contextual factors:

- speaking rate (fast, average and slow at sentence level)
- vowel stress (unstressed, primary, and secondary lexical stress as found in the dictionary)
- location of the syllable in the word (final, penultimate, other, monosyllable)
- location of the syllable in the utterance (final, penultimate, and other position)

For instance, for the long vowel /i/ as in ‘beat’, the overall average duration over all 4,626 occurrences in the training set is 95 ms, with a standard deviation of 39 ms. However for fast-rate unstressed realizations (796 occurrences) its mean duration is only 78 ms (sd=25 ms), whereas for an average-rate, word-final and utterance-final position (12 occurrences) the average vowel duration is not less than 186 ms (sd=52 ms), see Fig. 3.

Selfloops in a multi-state Markov model are certainly able to capture part of this variability, but given its systematic

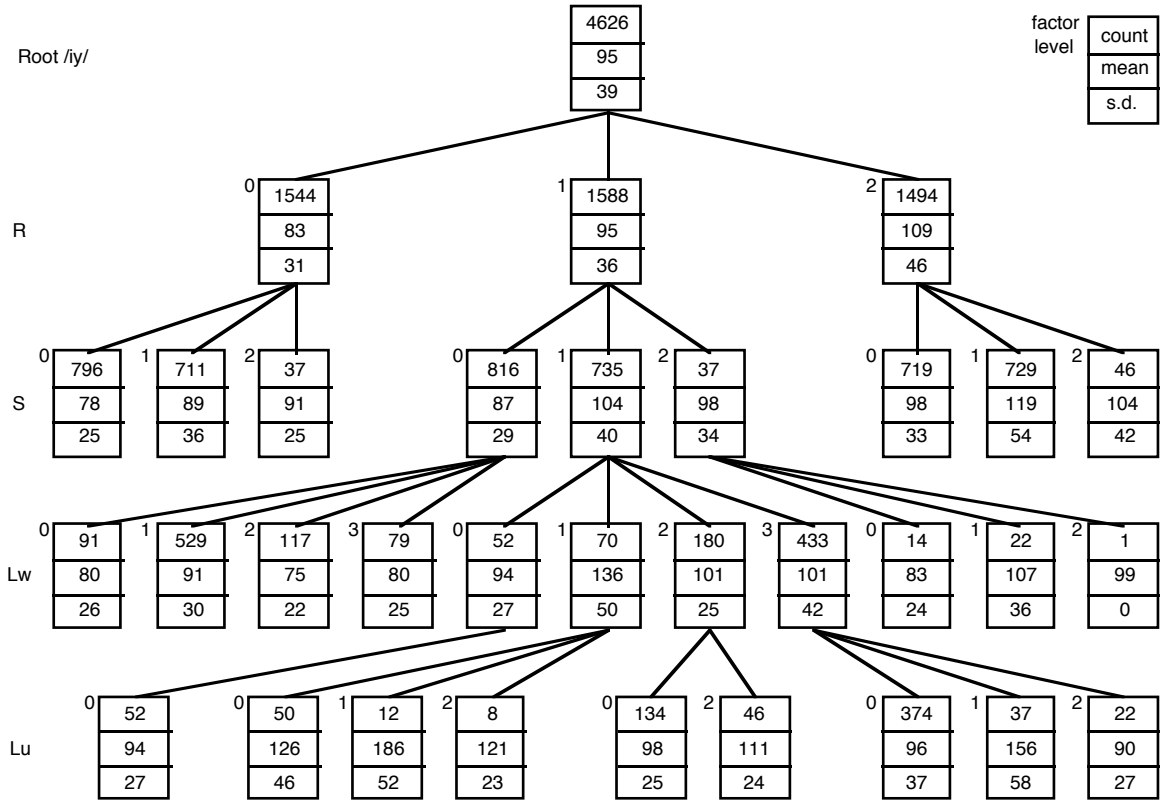

Figure 3. Part of the duration distribution of all 4,626 vowel /i/ segments in the TIMIT training set. Counts are number of phone instances per factor level. Mean duration and standard deviation are given in milliseconds. The factors are speaking rate *R*, at 3 levels (0=fast, 1=average, 2=slow), vowel stress *S*, at 3 levels (0=unstressed, 1=primary, 2=secondary), location of syllable in the word *Lw*, at 4 levels (0=other, 1=final, 2=penultimate, 3=mono), and location of syllable in the utterance *Lu*, at 3 levels (0=other, 1=final, 2=penultimate) [53].

behavior, one wonders whether a more condition-specific description could not be more helpful. Within the standard HMM toolkit that we had available, together with a rather simple N-best recognizer, we could only show marginal improvement from the base-line scores. I am convinced that an integrated approach would give additional progress.

This type of durational information [42] is certainly most useful in rule-based synthesis, especially since there one prototype is good enough, whereas in recognition one always has to worry about individual variability.

5.2 Vowel and consonant reduction, coarticulation

Similarly, phoneticians wonder whether specific spectral information could improve recognition. Spectral variability is not random, but at least partly speaker-, style-, and context-specific: small-headed speakers have higher formants than big-headed ones, schwa realization is not a simple centralization process but is strongly controlled by local context, fast and sloppy pronunciation shows more reduction than hyper speech, liquids and nasals do something to vowel quality, new and thus generally stressed information is more clearly articulated than given information.

Again, the observation probabilities in a Markov model can take care of a lot of spectral variability, especially so when multiphone-models are used, however, whenever such variability is systematic, it might still be worthwhile to model that. So, why not have separate models for full and reduced vowels? Not even in diphone synthesis it is very common to have at least two diphone sets, one for full and one for reduced vowels. Sometimes, system designers are lucky while they get spectral reduction for free in shortening the segment.

Why not distinguish between stressed and unstressed, and why not between read and spontaneous speech? Most people will take for granted that there are consistent distinctions between these conditions for *vowels*. Van Son & Pols [47] showed that this is similarly true for *consonants*. Acoustic consonant reduction can be expressed in terms of such measures as:

- duration
- spectral balance
- intervocalic sound energy differences
- F2 slope difference
- locus equation

Fig. 4 gives an example of the overall results, here on mean consonant duration, split on speaking style and syllable stress for 791 VCV segments taken from spontaneous and corresponding read speech from a single male Dutch speaker. Differences between conditions are substantial and indicate consonant reduction in spontaneous speech and in unstressed segments.

These results correlate nicely with consonant identification results in a listening experiment with 22 Dutch subjects using the same VCV stimuli [47], see Fig. 5.

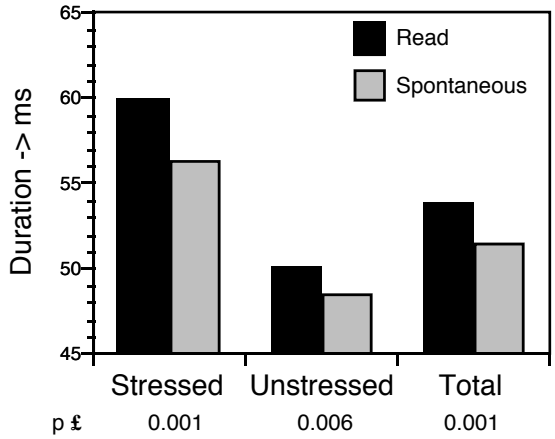

Figure 4. Mean durations (in ms) of the consonant tokens, split on speaking style (read and spontaneous) and syllable stress. The significance levels of the differences between read and spontaneous realizations are calculated using the Wilcoxon Matched-Pairs Signed-Ranks test.

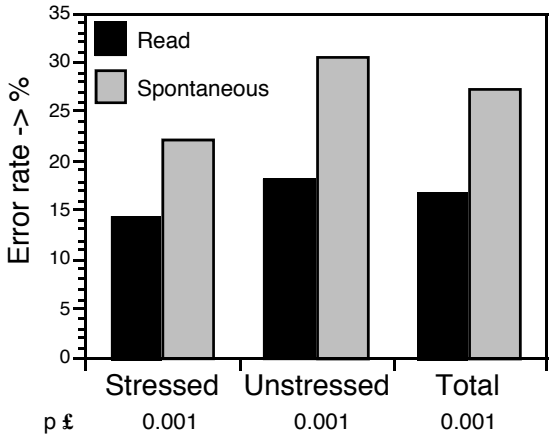

Figure 5. Mean error rates for consonant identification, split on speaking style (read and spontaneous) and syllable stress. The significance levels of the differences between read and spontaneous realizations are calculated using McNemar's test.

5.3 Pronunciation variation

Most recognizers work with a lexicon in which all words in the vocabulary have their normative pronunciation. Everybody knows that actual pronunciation can deviate substantially from that norm [52], see also sect. 5.2 and 5.5. Again, skips in a Markov model are quite powerful in modeling potential deletion and the like in a probabilistic way. However, certain substitution, reduction and deletion phenomena are much more systematic, and could perhaps become part of the sequential word model itself.

5.4 Speech efficiency

Recently we started a new project on the efficiency of speech [46]. Speaking is considered to be *efficient* if the speech sound contains *only* the information needed to understand it. This was expressed nicely by Lindblom [23] in saying *'speech is the missing information'*. On a corpus of spontaneous and corresponding read speech we indeed found that the duration and spectral reduction of consonants and correlate with the syllable and word frequency in this corpus. Consonant intelligibility in VCV segments correlates with both the acoustic factors and the syllable and word frequencies. It might be interesting in future recognizers to integrate this statistical knowledge with acoustic and n-gram language knowledge.

5.5 Units in speech recognition

Greenberg [13] presents very interesting data about a detailed analysis of 4 hours of phonetically labeled data of the Switchboard corpus (informal, unscripted, telephone dialogs in American English). The 100 most common words account for 66% of all individual tokens (25,923). The 30 most frequent words are all monosyllabic, whereas from the next 70 words only 10 are not. Eighty one percent of all corpus tokens are monosyllabic although they cover only 22.4% of the word types. The most common words, such as 'I', 'and', 'the', 'you', 'that', 'a', 'to', 'know', 'of', and 'it', all show substantial variation in pronunciation. On average there are 62 different phonetic expressions per word! Jurafsky et al. [19] indicate systematicity in the amount of reduction in these 10 function words. Also for most other words the phonetic realization in spontaneous speech often differs markedly from the canonical, phonological representation. According to Greenberg the patterns of deletion and substitution become rather systematic when placed within the framework of the syllable. He concludes that the syllable really is the basis for pronunciation and could profitably be used as the basis for recognition as well.

5.6 Quick adaptation

One of the most intriguing capabilities of human listeners, is their quick adapatation to new speakers, speaking styles, and environmental conditions. Probably most astonishing of all is the child's capability to understand her mother's and even her father's speech, despite substantial differences with her own speech. Various speaker normalizations in the vowel formant space have been proposed over the years, but none is really effective or appealing, partly because additional knowledge is required. Perception experiments with blocked-speakers- and mixed-speakers-designs have given some insight, but are rather artificial [3]. There is this tradeoff between quick adaptation, continued learning and buffering of old memories, that Grossberg [14] calls the stability-plasticity dilemma. He proposes the Adaptive Resonance Theory (ART) as one of the solutions to that problem.

In most speech recognizers the input variability is either incorporated in the training data, which is a rather brute force approach, or some form of adaptation is applied. One way is hierarchical codebook adaptation [12], but also tree-based dependence models are getting popular now [21]. Approaches used to personalize a synthetic voice [20] may also be interesting.

5.7 Prosody-driven recognition

Prosody mainly shows itself in accentuation and boundary marking. It provides important cues about word stress and sentence accent, and thus about given and new information. Durational and intonational characteristics mark phrase boundaries. Prosody provides important communicative information that is indispensable for text interpretation [8, 9] and dialog disambiguation [27]. Nevertheless it is barely used so far in ASR for several reasons. Prosody is a supra-segmental feature and thus difficult to handle by frame-based recognition systems.

Furthermore, error-free pitch extractors, working directly on the microphone signal, do not yet exist, whereas also a proper interpretation of the raw F0-contour is not an easy task. Even if segments are properly located, their duration cannot so easily be interpreted in a relative way. For instance, phrase-final lengthening is a nice concept, but the occurrence of a long syllable has to be detected relative to local speaking rate, the actual phonemes in the syllable, the length of the word, etc..

We are presently running a project about finding acoustic correlates for prominence, in which we envisage many of the above problems. Most naive language users don't know about metrical phonology, new/given, accent-lending pitch movements, break indices, and the like. However they can mark the word(s) in a sentence that they perceive as being spoken with prominence. We would like to find the best set of acoustic features and the best algorithm to predict perceived prominence directly from the speech signal. Using F0-range and duration per syllable, as well as loudness per vowel, as prominence predictors shows promising results [51], but more detailed information is needed [17, 38].

5.8 Multiple features

In my opinion, one of the main differences between human and machine speech processing is the fact that humans use multiple sources of information and select from them upon demand [44], whereas machines are operating with a fixed set of features and fixed procedures for recognition. Disambiguating between two minimally different words requires another level of spectro-temporal resolution than speech-non-speech detection. Using the appropriate spectral and temporal selectivity, preferably from parallel channels in which all varieties are available for some time, plus optimal use of multiple cues and trade-off relations [26, 39] is characteristic for efficient human performance.

6. DISCUSSION

In the above presentation I have stressed once again the well-known fact that humans generally do much better than machines in recognizing speech. I also tried to indicate how and why humans frequently do better. However, most of the time it was not so easy to conclude what knowledge, so far, was neglected in ASR, how that easily could be added, and what specific increase in performance that would bring under certain conditions. Please don't blame me for that. I am simply a scientist with a background in phonetics and speech perception who has a strong interest in speech technology and who seriously believes that substantial progress in speech technology still can be made by learning from human functionality. Of course, improving the predictability of communication by proper dialog handling and language modeling, will be extremely helpful, but still also much progress can be gained from optimal front-end processing and acoustic modeling and recognition.

I also want to repeat the pledge made by Roger Moore [25] at ICPhS'95 in Stockholm for *Computational phonetics*. He then indicated that

> *"the skills and expertise represented by the phonetic science community could be usefully directed **not** towards the construction of better automatic speech recognisers or synthesisers, but towards the exploitation of the theoretical and practical tools and techniques from speech technology for the creation of more advanced theories of speech perception and production (by humans and by machines)".*

It will hopefully be clear that I believe that both communities could benefit from each other.

I do believe that (computational) phonetics does make interesting contributions to speech technology via duration modeling (e.g., [41, 49, 53], pronunciation variation modeling [52], vowel reduction models (e.g., [2]), computational prosody [40], using appropriate information measures for confusions [45], the use of the modulation transfer function for speech [18], etc..

Similarly does speech technology, with its abundance of (automatically or hand-annotated) speech and language databases and its wealth of analysis, modeling, and recognition methods, provide interesting tools to collect speech data, to extract regularities, and to apply that knowledge.

REFERENCES

[1] Benoit, C., Grice, M. & Hazan, V. (1996), "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences", *Speech Communication,* 18, 381-392.

[2] Bergem, D.R. van (1995), "Experimental evidence for a comprehensive theory of vowel reduction", *Speech Communication,* 16, 329-358.

[3] Bergem, D.R. van, Pols, L.C.W. & Koopmans-van Beinum, F.J. (1988), "Perceptual normalization of the vowels of a man and a child in various contexts", *Speech Communication,* 7, 1-120.

[4] Bezooijen, R. van & Jongenburger, W. (1993), "Evaluation of an electronic newspaper for the blind in the Netherlands", In: *Proc. ESCA Workshop on Speech and Language Technology for Disabled Persons,* Stockholm, 195-198.

[5] Campbell, W.N. (1997), "Synthesizing spontaneous speech", In: Sagisaka et al. (Eds.), 165-186.

[6] Cutler, A., Dahan, D. & Donselaar, W. van (1997), 'Prosody in the comprehension of spoken language: A literature review*", Language and Speech,* 40, 141-201.

[7] Doddington, G.R. (1995), "Spoken language technology discussion", *Proc. Spoken Language Systems Technology Workshop*, Austin, Morgan Kaufman Publ., Inc, 289-294

[8] Donzel, M.E. van (1999), "Prosodic characteristics of information structure in spontaneous discourse in Dutch", *Proc. ICPhS'99,* San Francisco.

[9] Donzel, M.E. van, Koopmans-van Beinum, F.J. & Pols, L.C.W. (1998), "Speaker strategies in the use of prosodic means in spontaneous discourse in Dutch", *Proc. ESCA Worskhop on Sound Patterns in Spontaneous Speech,* Aix-en-Provence, 135-138.

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

[11] 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-recption threshold of sentences*", J. Acoust. Soc. Am.,* 81, 465-469.

[12] Furui, S. (1992), "Speaker-independent and speaker-adaptive recognition techniques", In: S. Furui & M.M. Sondhi (Eds.)*, Advances in speech signal processing,* New York: Marcel Dekker, Inc., 597-622.

[13] Greenberg, S. (1998), "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation", In: Strik et al. (Eds.), 47-56.

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

[15] Hermansky, H. (1998), "Should recognizers have ears?", *Speech Communication,* 25, 3-27.

[16] Hermansky, H. & Sharma, S. (1998), "TRAPS - Classifiers of temporal patterns", *Proc. ICSLP'98,* Sydney, Vol. 3, 1003-1006.

[17] Hess, W., Batliner, A., Kiessling, A., Kompe, R., Nöth, Petzold, A., Reyelt, M. & Strom, V. (1997), "Prosodic modules for speech recognition and understanding in VERBMOBIL", In: Sagisaka et al. (Eds.), 361-382
[18] 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.
[19] Jurafsky, D., Bell, A., Fosler-Lussier, E, Girand, C. & Raymond, W. (1998), "Reduction of English function words in Switchboard", *Proc. ICSLP'98*, Sydney, Vol. 7, 3111-3114.
[20] Kain, A. & Macon, M. (1998), "Personalizing a speech synthesizer by voice adaptation", *Proc. ESCA Workshop on Speech Synthesis*, Jenolan Caves, 225-230
[21] Kannan, A. & Ostendorf, M. (1997), "Modeling dependency in adaptation of acoustic models using multiscale tree processes", *Proc. Eurospeech'97,* Rhodes, Vol. 4, 1863-1866.
[22] 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.
[23] Lindblom, "Role of articulation in speech perception: Clues from production", *J. Acoust. Soc. Am.,* 99, 1683-1692.
[24] Lippmann, R.P. (1997), "Speech recognition by machines and humans", *Speech Communication,* 22, 1-15.
[25] Moore, R.K. (1995), "Computational phonetics", *Proc. ICPhS'95*, Stockholm, Vol. 4, 68-71.
[26] Nearey, T.M. (1997), "Speech perception as pattern recognition", *J. Acoust. Soc. Am.,* 101, 3241-3254.
[27] Nöth, E., Batliner, A., Kie**b**ling, A., Kompe, R. & Niemann, H. (1998), "Suprasegmental modeling", In: K. Ponting (Ed.), *Computational models of speech pattern processing*, Berlin: Springer Verlag, 182-199.
[28] Oostdijk, N., Goedertier, W. & Martens, J.-P. (1999), "The Spoken Dutch Corpus Project", *Proc. Eurospeech'99,* Budapest.
[29] Pallett, D.S. (1998), "The NIST role in automatic speech recognition benchmark tests", *Proc. LREC'98*, Granada, Vol. 1, 327-330.
[30] Pallett, D.S., Fiscus, J.G., Fisher, W.M., Garofolo, J.S., Lund, B.A., Martin, A. & Przybocki, M.A. (1995), "1994 Benchmark tests for the ARPA spoken language program", *Proc. Spoken Language Systems Technology Workshop*, Austin, Morgan Kaufman Publ., Inc, 5-36.
[31] Pols, L.C.W. (1981), "Consonant intelligibility in reverberant and/or ambient noise conditions", *Proc. 4th FASE Symp. on Acoustics and Speech,* Venice, 87-90.
[32] Pols, L.C.W. (1982), "How humans perform on a connected-digits data base", *Proc. IEEE-ICASSP'82,* Paris, Vol. 2, 867-870.
[33] Pols, L.C.W. (1997), "Flexible human speech recognition", *Proc. ASRU'97*, Santa Barbara, Piscataway, NJ: IEEE Signal Processing Society, 273-283.
[34] Pols, L.C.W. (1998a), "Foreword", In: R. Sproat (Ed.), *Multilingual text-to-speech synthesis. The Bell Labs approach*, Dordrecht: Kluwer Academic Publishers, xxiii-xxiv.
[35] Pols, L.C.W. (1998b), "Psycho-acoustics and speech perception", In: K. Ponting (Ed.), *Computational models of speech pattern processing*, Berlin: Springer Verlag, 10-17.
[36] Pols, L.C.W., Santen, J.P.H. van, Abe, M., Kahn, D. & Keller, E. (1998), "The use of large text corpora for evaluating text-to-speech systems", *Proc. LREC'98*, Granada, Vol. 1, 637-640.
[37] Pols, L.C.W., Wang, X. & Bosch, L.F.M. ten (1996), "Modeling of phone duration (using the TIMIT database) and its potential benefit for ASR", *Speech Communication,* 19, 161-176.
[38] Portele, T. (1998), "Perceived prominence and acoustic parameters in American English", *Proc. ICSLP'98*, Sydney, Vol. 3, 667-670.
[39] Repp, B. (1982), "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception", *Psychological Bulletin,* 92, 81-110.
[40] Sagisaka, Y, Campbell, N. & Higuchi, N. (Eds.) (1997), *Computing prosody. Computational models for processing spontaneous speech,* New York: Springer, 401 pp.
[41] Santen, J.P.H. van (1992), "Contextual effects on vowel duration", *Speech Communication,* 11, 513-546.
[42] Santen, J.P.H. van (1997), "Prosodic modeling in text-to-speech synthesis", *Proc. Eurospeech'97*, Rhodes, Vol. 1, KN-19-28.
[43] Santen, J.P.H. van, Pols, L.C.W., Abe, M., Kahn, D., Keller, E. & Vonwiller, J. (1998), "Report on the third ESCA TTS workshop evaluation procedure", *Proc. ESCA Workshop on Speech Synthesis*, Jenolan Caves, 329-332.
[44] Smits, R. (1996), "A pattern-recognition-based framework for research on phonetic perception*", Speech Hearing and Language: work in progress,* 9, 195-229.
[45] Son, R.J.J.H. van (1995), "A method to quantify the error distribution in confusion matrices", *Proc. Eurospeech'95,* Madrid, Vol. 3, 2277-2280.
[46] Son, R.J.J.H. van, Koopmans-van Beinum, F.J. & Pols, L.C.W. (1998), "Efficiency as an organizing principle of natural speech", *Proc. ICSLP'98*, Sydney, Vol. 6, 2375-2382.
[47] 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.
[48] Son, R.J.J.H. van & Pols, L.C.W. (1999), "An acoustic description of consonant reduction", *Speech Communication* (accepted for publication).
[49] Son, R.J.J.H. van & Santen, J.P.H. van (1997), "Strong interaction between factors influencing consonant duration*", Proc. Euro-speech'97,* Vol. 1, 319-322.
[50] Steeneken, H.J.M. (1992*), On measuring and predicting speech intelligibility,* Ph.D thesis, University of Amsterdam, 165 pp.
[51] Streefkerk, B.M., Pols, L.C.W. & Bosch, L.F.M. ten (1999), "Towards finding optimal acoustical features as predictors for prominence in read aloud Dutch sentences", *Proc. ICPhS'99,* San Francisco.
[52] Strik, H., Kessens, J.M. & Wester, M. (Eds.) (1998), *Proc. ESCA Worskhop on Modeling pronunciation variation for automatic speech recognition*, Rolduc, 162 pp.
[53] Wang, X. (1997), *Incorporating knowledge on segmental duration in HMM-based continuous speech recognition*, Ph.D. thesis, University of Amsterdam, SLLU 29, 190 pp.
[54] Wieringen, A. van (1995), *Perceiving dynamic speechlike sounds. Psycho-acoustics and speech perception,* Ph.D. thesis, University of Amsterdam, 256 pp.
[55] 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, 1304-1312.
[56] Zue, V. (1985), "The use of speech knowledge in automatic speech recognition", *Proc. of the IEEE,* 73(11), 1602-1615.
[57] Zue, V. & Seneff, S. (1990), "Transcription and alignment of the Timit database", In: H. Fujisaki (Ed.), *Recent research toward advanced man-machine interface through spoken language*, Tokyo, 464-473.