STUDYING SPEECH PERCEPTION: FROM SIMPLE STATIONARY STIMULI TO MORE AND MORE COMPLEX SPEECH(-LIKE) SIGNALS [1]

Louis C.W. Pols

ABSTRACT

In psycho-acoustic research it is common practice to use rather simple, mainly stationary, *stimulus sounds* such as pure tones. However, concerning the *research methods* the approaches chosen are frequently very advanced. As a contrast, in most speech perception research the (speech) stimuli are dynamic in nature and spectrally complex, but the task is often a simple forced-choice identification. It might be very useful to combine the best of both approaches, in that way gaining more insight in the process of natural speech perception as well as in the understanding of synthetic speech.

1. INTRODUCTION

In order to understand better the characteristics of speech as well as the process of speech perception and understanding, the (speech) *signal* is frequently simplified and/or the *perception task* is reduced to one specific aspect only.

Simplifying the *signal* makes it possible to study the importance of one or a few characteristics on its own merit, such as the importance of formant positions for vowel identification. However, even a naturally spoken isolated vowel sometimes is already too complex as a basis for studying specific aspects of vowel perception since, apart from the inherent spectral variation from one vowel to the other, many other characteristics will vary simultaneously such as fundamental frequency, overall energy, duration, dynamic spectral variation, source characteristics, background noise, etc. Full control of all parameters seems only to be possible by using synthetic speech. The inherent danger of this approach apparently lies in the potentially big gap between well-defined, but artificial, synthetic speech and ill-defined, but natural, real speech.

Reducing the *task* of the subject to one specific aspect permits isolation of that component. Examples are, for instance, the task of judging the loudness or the duration while neglecting other aspects of the signal, or other tasks such as vowel identification, or F2' matching.

There are many parallels between, on the one hand, psycho-physical research of sound and of music perception and, on the other hand, phonetic and psychologic research of speech perception. This holds just as much for the signals used as for the methods applied.

[1] Written version of a paper presented at the 28th Acoustic Conference, 3-6 October 1989, Štrebské Pleso, Czechoslowakia.

The basic perceptual concepts of pitch, loudness, duration, and timbre are equally valid for speech, although the human origin, the dynamic variation, and the linguistic content make speech different from other sounds.

Psycho-physical methods, such as signal detection, masking, identification, forced-choice judgments, and matching are all potentially useful in speech research too.

Below we will first present a scheme to classify speech stimuli, next we will present in an ordered way the many different methods for doing perceptual experiments. In the subsequent section the two will be combined and examples will be given of listening experiments, mainly in the domain of diagnostic evaluation of the speech quality of rule-based text-to-speech synthesis systems.

2. CLASSIFICATION OF SPEECH SIGNALS AND PERCEPTION METHODS

In Table I below we present, from top to bottom, a continuum from fully natural speech down to very simple stationary speech-like sounds, and then up again to more-and-more-natural synthetic speech. For (manipulated) natural speech a number of examples are given of speech signals regularly used in speech research. In the lower half of the table only the various synthesis methods are mentioned. By applying these synthesis methods one can in principle generate any aforementioned speech signal, all the way from a single vowel period up to a paragraph of text, although one will have to take into account the unavoidable quality degradation.

In Table II below a number of psycho-acoustic, psycho-linguistic, and speech perception methods are presented, from signal detection and discrimination, via matching and identification, to evaluation.

Table I. Speech signal classification and possible synthesis methods.

Natural speech: free conversation, read text, sentences, words, syllables.

Manipulated natural speech: 100-msec vowel segments, one-period vowels, plosive bursts, vocalic transitions, intonation groups, reiterant speech, word segments, speech segment from one context presented in another context, controlled envelope, transformed speech, whispered speech, bite-block speech, masked or filtered speech.

Controlled speech-like sounds: Pattern Playback ba-da-ga, harmonic summation, n-formant vowels, F1-F2', inverted speech.

Speech based on analysis/resynthesis: diphone speech, allophone-based speech, articulatory synthesis, duration rules, intonation contours, sinus speech.

Speech from synthesis by rule: from concept or from text.

Table II. Psycho-physical, psycho-linguistic, and speech perception methods.

- *detection*: absolute vs. masked threshold (in noise)
- just noticeable difference (JND)
- difference limen (DL)
- forward/backward masking
- gap detection, modulation detection
- continuity threshold (internal spectrum)

- *discrimination*: ABX, same-different judgment (AA - AB),
  two-alternatives-forced-choice (2AFC),
  similarity judgments in pairs or triads

- *matching* a controllable signal with a test signal

- *identification*: (open/closed response) with(out) distortion,
  selective adaptation,
  categorical perception
- (narrow) transcription
- gating paradigm
- phoneme monitoring, shadowing

- *memory recall* in (out of) order
- lexical decision (word vs. non-word)

- *choice reaction time*

- *evaluation* by semantic scaling (on bipolar 7-point scales)
- preference judgment

3. VARIETY OF SPEECH SIGNALS AND PERCEPTION METHODS, AS USED IN SPEECH SYNTHESIS ASSESSMENT

In discussing the process of speech perception and the signals and methods available for that, one would probably not immediately consider the assessment of the quality of rule-synthesized speech a very relevant task (Pols, 1989). However, rule-synthesized speech has the great advantage that both the signal can be fully controlled (from the characteristics of a single sound to a full read-aloud paragraph), as well as the evaluation itself can cover a great range of tasks (from simple phoneme identification to memory recall and text comprehension). So it illustrates the usefulness of speech synthesis for a better understanding of the process of speech perception. Below I will limit myself to a few examples of phoneme intelligibility (section 3.1), then I will jump to a totally different aspect of synthetic speech, namely the prosodic characteristics and some ways to evaluate these (section 3.2). In the last section (3.3) a few other evaluation methods will just be mentioned. In chapter 4 some examples are given of 'alternative' ways to use a speech synthesizer, namely as a stimulus generator for studying basic speech sound characteristics and the perception of these.

3.1 Phoneme and word perception

Because of the voiced nature of *vowels*, with only gradual changes in the signal, most present-day synthesizers will have no great difficulty to generate acceptable vowel sounds, whether they are allophone-, diphone-, or syllable-based and whether they use LPC parameters or formants. In certain tests, such as the diagnostic rhyme test (DRT) (Voiers, 1977) or the modified rhyme test (MRT) (House et al., 1964), the vowel intelligibility is not even evaluated. If measured nevertheless, such as in a CVC test, whether or not with the phonemes phonetically balanced per list, one achieves correct vowel scores of 70% and beyond. Naturally, this depends somewhat upon the characteristics of the language and of the system. For example, the nasalized vowels in French are not so easy to generate with an all-pole LPC-model. In Italian there are effectively only 5 different vowels, which substantially reduces chances for any vowel error, Dutch counts 15 different vowels, and the Swedish language even more.

The *consonant* intelligibility is generally substantially worse, partly because of the abrupt nature of certain sounds, or its mixed-excitation, or its peculiar spectro-temporal characteristics in the transition region, or for whatever other reasons. Scores appear to differ also substantially depending upon the open or closed nature of the allowable response set. In my view it is better to use an open response task for diagnostic purposes, whereas a closed response set (such as used in DRT or MRT) makes some sense when evaluating the communicative power of a system. In natural speech there will always be a certain amount of redundancy in the text because of acoustic and linguistic context, and this redundancy can be simulated by presenting only certain response alternatives. Unfortunately, these alternatives are based on old studies from the 'analog world', using noise masking and bandpass filtering, whereas present-day digital synthesizers sometimes produce very peculiar defects not necessarily covered in the presented response alternatives.

A similar problem arises when studying in a diagnostic way phoneme intelligibility by presenting *words*. If, as is done frequently, these words are meaningful, then the response alternatives for the listener are more determined by word alternatives than by acoustic similarity at the phoneme level. I prefer for diagnostic tests nonsense words of the CVC- or VCV-type. This then indeed requires some training from the subjects. If the synthesis system is diphone-based it makes sense to use also CVVC- and VCCV-type words since otherwise the VV- and CC-diphones will not be evaluated at all (van Bezooijen & Pols, 1987; Pols et al., 1987; van Son et al., 1988).

A next level of evaluation concerns *consonant clusters*. So far I only know of few tests in which this topic is covered. In a Dutch test only initial and final clusters have been studied, whereas in a similar experiment for French also the word internal medial clusters were part of the test. Spiegel et al. (1989) included words with consonant clusters in his monosyllabic test corpus to evaluate the intelligibility of synthetized and natural American English speech (over the telephone).

3.2 Prosody and stress

With the word material discussed above, one actually only studies stressed syllables. However, in natural speech it is very important to be able to produce the right word stress at the right syllable, and give the other syllables secondary or no stress. This type of information, necessary for high-quality speech synthesis, could be stored in a word lexicon, although a set of rules would make it more universal. But even if one knew exactly where to put the word stress, it is not very clear how to put it there, and what to do with the remaining syllables. As far as I know, no methods have been proposed so far to evaluate the intelligibility of *unstressed syllables*. One might be able to generate a limited number of minimal pairs, such as *sub*ject and sub*ject*, but this can only be a partial solution. Going from words in isolation to words in sentences, the matter is further complicated by between-word boundary effects, such as assimilation, elision, and coarticulation.

Already in a multi-syllabic word, but even more so in sentences, the *prosodic features* are of utmost importance for a natural speech quality. To produce the right amplitude envelope, the right segmental duration and speaking rate, and the right intonation contour, with appropriate sentence accent, is one of the most difficult tasks in rule synthesis. However, in the present paper the question how to evaluate the perceived quality of such characteristics, is of greater concern to us. Relatively little is known about JNDs for pitch, duration, and amplitude (changes) in speech. The characteristics of high-pitched voices are even less well understood (see vanWieringen and Pols, this volume). The evaluation of prosodic characteristics of synthesized speech is generally limited to paired comparison, semantic scaling, or expert judgments.
In the ESPRIT-SAM project a universal multi-lingual structure has been developed for semantically anomalous sentences using meaningful words (Benoit et al., 1989). These sentences can be very useful for suprasegmental tests both with respect to intelligibility and prosodic evaluation.

Van Bezooijen performed an interesting experiment to evaluate the performance of an algorithm for defining the *sentence accent* in any Dutch text (van Bezooijen and Pols, 1989). She asked subjects to judge on a 10-points scale the adequacy of the rule-based as well as other accent structures, both from sentences on paper (with the accented words in CAPITALS) and from acoustic (diphone) realizations. The rule-based accentuation was judged slightly worse than resynthesized natural accents, but substantially better than semi-random accents. In the acoustic realizations all *pauses* were hand-marked and produced with equal length. This is another parameter to be optimized. Nooteboom (1983) showed that, especially for not too good a quality of synthetic speech, it can be beneficial to introduce (additional) pauses at the right places to relief the listeners temporarily from some form of processing overload.
Nusbaum & Pisoni (1985) showed that indeed listening to synthetic speech of a less-than-optimal quality requires extra *processing time*.

3.3 Other perception methods used in synthesis evalation

In describing above some of the various levels of synthesis evaluation, we have already come across several of the methods presented in Table II. Some other examples are the word gating experiment performed by Nooteboom and Doodeman (1983) using diphone speech, or various other psycholinguistic experiments performed by Pisoni (1982), using lexical decision, phoneme monitoring, choice reaction times, memory recall, etc.

4. PERCEPTION OF BASIC SOUND CHARACTERISTICS

Once a scientist has access to the internal control mechanism of a rule synthesizer, he actually has also access to an advanced (speech) signal generator with which many interesting speech-like signals can be generated in order to study aspects of speech perception. Let me just give a few examples:

- Isolated tone, band and formant sweeps can be generated and presented for identification, discrimination (Pols and Schouten, 1987), or matching (Pols et al., 1984) in order to study aspects of the perception of dynamic formant transitions.
- Broadening of the auditory filters, as may be one of the characteristics of hearing impairment, can be simulated by manipulating the formant bandwidth of the stimuli used (Dubno and Dorman, 1987).
- The perceptual relevance of spectro-temporal characteristics related to change in, for instance, speaking rate, stress, coarticulation, and reduction can be studied by systematically manipulating the stimuli, and/or by varying the context in which they are presented to subjects.

5. CONCLUSIONS

Present-day speech technology virtually permits the generation of any desired signal characteristic. However, the actual quality of nowadays rule synthesizers is far from perfect, indicating that the rules governing the (systematic) variation in the speech signal, as controlled by language requirements and external conditions such as speaker, style, and communication channel, are far from understood. However, the same speech technology allows us to have good control over our speech stimuli, and allows us to run advanced and sophisticated listening experiments. We just have to come up with the right questions and the right ways to study and answer them!

REFERENCES

Benoit, C., Erp, A. van, Grice, M., Hazan, V. & Jekosch, U. (1989), Multilingual synthesiser assessment using semantically unpredictable sentences", Proc. EUROSPEECH'89, Paris, Vol. 2, 633-636.

Bezooijen, R. van & Pols, L.C.W. (1987), "Evaluation of two synthesis-by-rule systems for Dutch", Proc. Eur. Conf. Speech Techn., Edinburgh, Vol. 1, 183-186.

Bezooijen, R. van & Pols, L.C.W. (1989), "Evaluation of a sentence accentuation algorithm for a Dutch text-to-speech system", Proc. EUROSPEECH'89, Paris, Vol. 1, 218-221.

Dubno, J.R. & Dorman, M.F. (1987), "Effects of spectral sharpening on vowel identification", J. Acoust. Soc. Amer. 82(5), 1503-1511.

House, A.S., Williams, C.E., Hecker, M.H.L. & Kryter, K.D. (1964), "Articulation-testing methods: Consonantal differentiation with a closed-response set", J. Acoust. Soc. Amer. 37, 158-166.

Klatt, D.H. (1987), "Review of text-to-speech conversion for English", J. Acoust. Soc. Amer. 82, 737-793.

Nooteboom, S.G. (1983), "The temporal organization of speech and the process of spoken-word recognition", IPO Annual Progress Report 18, 32-36.

Nooteboom, S.G. & Doodeman, G.J.N. (1982), "Speech quality and word recognition from fragments of spoken words", IPO Annual Progress Report 17, 46-56.

Nusbaum, H.C. & Pisoni, D.B. (1985), "Constraints on the perception of synthetic speech generated by rule", Behavior Research Methods, Instruments, & Computers 17(2), 235-242.

Pisoni, D.B. (1982), "Perception of speech: The human listener as a cognitive interface", Speech Technology 1-2, 10-23.

Pols, L.C.W. (1989), "Assessment of text-to-speech synthesis systems", In: A. Fourcin et al. (Eds.) Speech input and output assessment. Multilingual methods and standards, Ellis Horwood Limited, Chichester, 53-81 & 251-266.

Pols, L.C.W., Boxelaar, G.W. & Koopmans-van Beinum, F.J. (1984), "Study of the role of formant transitions in vowel recognition using the matching paradigm", Proc. Inst. of Acoustics, Vol. 6, Part 4, 371-378.

Pols, L.C.W., Lefevre, J.P., Boxelaar, G. & Son, N. van (1987), "Word intelligibility of a rule synthesis system for French", Proc. Eur. Conf. Speech Techn., Edinburgh, Vol. 1, 179-182.

Pols, L.C.W. & Schouten, M.E.H. (1987), "Perception of tone, band, and formant sweeps", In: M.E.H. Schouten (Ed.), The psychophysics of speech perception, Martinus Nijhoff Publishers, Dordrecht, 231-240.

Son, N. van, Pols, L.C.W., Sandri, S. & Salza, P.L. (1988), "First quality evaluation of a diphone-based speech synthesis system for Italian", Proc. SPEECH'88, 7th FASE Symp., Edinburgh, Book 2, 677-682.

Spiegel, M., Altom, M.J., Macchi, M. & Wallace, K. (1988), "A monosyllabic test corpus to evaluate the intelligibility of synthesized and natural speech", Proc. ESCA Workshop on 'Speech input/output assessment and speech databases', Noordwijkerhout, The Netherlands, 1.2.1-1.2.5

Voiers, W.D. (1977), "Diagnostic evaluation of speech intelligibility", In: M.E. Hawley (Ed.), Speech intelligibility and speaker recognition, Dowden, Hutchinson, and Ross, Stroudsberg, PA, 374-387.