Proceedings of The Institute of Acoustics

PERCEPTION OF PROCESSED SPEECH AND SOME IMPLICATIONS FOR ENHANCEMENT

P J BLOOM

POLYTECHNIC OF CENTRAL LONDON

INTRODUCTION

The purpose of this paper is to suggest some plausible objectives for a speech enhancement system in the context of observed perceptual phenomena and models of speech perception. To do this we shall define the enhancement problem in relation to perception, present a few examples of relevant, speech-based perceptual phenomena, and outline one intriguing model of speech perception which may contain useful hints for perceptually based signal processsing. It is hoped that this informal presentation may stimulate further discussions and more formal investigations in this area.

THE RELATIONSHIP OF SIGNAL PROCESSING TO PERCEPTUAL MODELS

A very simplified view of the lowest levels of speech perception, i.e. the acoustic and phonetic levels, is presented in Fig. 1a. There, a time-varying speech signal, $s(t)$, is transformed in a generalised model of speech perception into the percepts of a set of phonemes, $\{P_1(t)\}$. Needless to say, a great deal is known about the relationships between the acoustic signal and the phonemes. In Fig. 1b, noise-corrupted speech enters the model and a second set of phonetic responses $\{P_2(t)\}$ emerges. Depending on the acoustic characteristics of the degraded signal, the phonemes may or may not be "perceived" correctly.

In this example one should be aware that no current perceptual model is able to accurately predict what phonemes an observer will hear. If, as shown in Fig. 1c, we introduce signal processing to generate an "enhanced" signal $\hat{s}(t)$ which contains elements of the original speech plus some possibly unwanted modifications, the situation becomes more complicated. The critical question is now whether the altered acoustical signal features in $\hat{s}(t)$ which produce the phonetic response $\{P_3(t)\}$ make the recognition task easier or more difficult. i.e. are the processed acoustical features now more or less suitable for the perceptual processing to follow. Another way to look at this is to ask what criteria should we apply to the signal processor in allowing some permissable level of error signal

$$e(t) = \hat{s}(t) - s(t)$$

to remain. Traditionally (and obviously most conveniently) the above is generally approached from a mathematical viewpoint: e.g. minimize the mean squared error. Additional constraints are generally imposed based on information assumed from fairly rigid mathematical models of the speech and noise waveforms or the speech production mechanisms.[1] However, we shall try to indicate in the next section that this approach may not always provide the optimum solution in the context of enhancing speech intelligibility and that criteria based on a good perceptual model would be more desirable.

PERCEPTION OF PROCESSED SPEECH AND SOME IMPLICATIONS FOR ENHANCEMENT

EXAMPLES OF ALTERED SPEECH PERCEPTION

In this section we discuss the perception of various computer processed signals in which the signal alterations are physically well-defined and the effects on corresponding streams of phonemes are also well known and demonstratable. From such relationships we can state some general requirements for a speech perception model.

A) Short Term Spectral Alterations: The short term amplitude spectrum carries the primary information for phonome identification. The fine structure of the speech waveform carries little phonetic information. This may be easily demonstrated by comparing a speech signal synthesized from normal speech whose original short term magnitude spectrum is retained and whose short term phase spectrum is random, with a signal whose original short time phase is retained and whose magnitude is kept constant. The former is intelligible (sounding like a hoarse whispered voice) but the latter contains only noise or pitch pulses. This result implies that a perceptual model must tolerate severe waveform distortion but not severe short term spectral magnitude distortion.

B) Silence vs. Noise-Filled Deleted Speech Intervals: A particularly compelling aspect of speech perception (that would not generally be predicted on the basis of mathematical considerations of signal waveforms) is that if segments of speech are deleted at certain rates, replacement of the deleted segments by wide-band noise can improve intelligibility considerably over the condition of silence in the deleted interval [2,3]. Moreover the measured improvement can be greatest for the lowest signal to noise ratios (i.e. highest noise levels) [2]. This effect was studied in detail using Dutch consonants and only deleted initial /t/,/p/,and /k/'s were positively somewhat restored by adding noise [3]. One suggested explanation is that noise tends to mask misleading cues present when speech is switched abruptly to silence [2], but it is not clear whether the transients or signal absence generates the misleading cues. Nevertheless, these results are important to incorporate into a useful perceptual model, inasmuch as they provide some indication of gross constraints on the levels of signal suppression permissable in speech processing systems.

A MODEL FOR SPEECH PERCEPTION

We lastly consider a rather intriguing theory of speech perception which was proposed by Yilmaz [4,5]. In this theory, he postulated a perceptual space of three main dimensions into which acoustic stimuli could be transformed into whispered (and voiced) vowels and consonants. In the first stage of his model, the physical variables of frequency and intensity are mapped into a psychophysical space, linear in mels and sones. The resulting time-varying mel spectral distribution, $S(p,t)$ where p is pitch and t time, may be synthesized by a linear combination of m basic mel spectral distributions (or principle component basis vectors [6]) which we denote by $b_o(p)$, $b_1(p),\ldots,b_m(p)$. Thus,

$$S(p,t) = w_o(t)b_o(p) + w_1(t)b_1(p) + \ldots + w_m(t)b_m(p)$$

and the coefficients w_o, w_1, ... , w_m are the time-varying weights of each corresponding basis vector. Using only three basis vectors whose "shapes"

roughly resemble a Fourier series set on a restricted-range mel scale transformed to degrees through the variable ϕ (i.e. $b_0 \cong 1$, $b_1(\phi) \cong \sin\phi$ and $b_2(\phi) \cong \cos\phi$), Yilmaz postulated a three dimensional space along w_0, w_1, and w_2 in which all phonemes are systematically distributed. Because b_0 is related to intensity, the different speech sounds are found to be distributed in a circle in the b_1 - b_2 plane, and the element of time primarily distinguishes vowels from consonants.

The reason this type of model may be of interest in speech processing (even though these exact transformations might not take place in human audition) is that it appears to provide a phonemic output that is relatively invariant, firstly under conditions of different speakers and varied prosodic features, and secondly, under noise conditions if an adaptive transformation of the vector space is effected. Some interesting experiments are described [4,5] based on this theory, some of which we shall demonstrate.

Such a model could be useful in two distinct ways. First, signal processing could be envisaged within the perceptual space that implements noise reduction and an inverse transformation could be applied to generate an acoustical signal. (A related approach was taken in [7] where noise suppression processing was implemented in the perceptual domains of critical band frequency analysis and loudness and was found to be effective.) Alternatively, this model could be used to "evaluate" processing systems, as in Fig.1c.

SUMMARY

In summary, we have tried to outline some objectives for speech enhancement systems in terms of general perceptual requriements, rather than in terms of the physical characteristics of the signal. In addition, we have suggested as a starting point for such work an interesting, yet unexplored model of speech perception.

ACKNOWLEDGEMENTS

This work was partially funded by the Science and Engineering Research Council, which the author greatfully acknowledges. The author also wishes to thank Prof. Cam Searle in the Reasearch Laboratory of Electronics of MIT for introducing him to the work of Yilmaz.

REFERENCES

1. J.S. Lim and A.V. Oppenheim 1979 Proc.IEEE 67(12) 1586-1604. Enhancement and Bandwidth Compression of Noisy Speech.

2. G. Powers and J. Wilcox 1977 J. Acoust.Soc.Am. 61(1) 195-199. Intelligibility of temporally interrupted speech with and without intervening noise.

3. L. Pols and M. Schouten 1978 J.Acoust.Soc.Am. 64(5) 1333-1337. Identification of deleted consonants.

PERCEPTION OF PROCESSED SPEECH AND SOME IMPLICATIONS FOR ENHANCEMENT

4. H. Yilmaz 1967 Bull.Math.Biophysics 29 793-825. A Theory of Speech Perception.

5. H. Yilmaz 1968 Bull.Math.Biophysics 30 455-479. A Theory of Speech Perception: II.

6. S.A. Zahorian and M. Rothenberg 1981 J.Acoust.Soc.Am. 69(3) 832-845. Principal-components analysis for low-redundancy encoding of speech spectra.

7. T.L. Petersen 1981 UTEC-CSc-80-113 Computer Science, Univ.of Utah. Acoustic Signal Processing in the Context of a Perceptual Model.

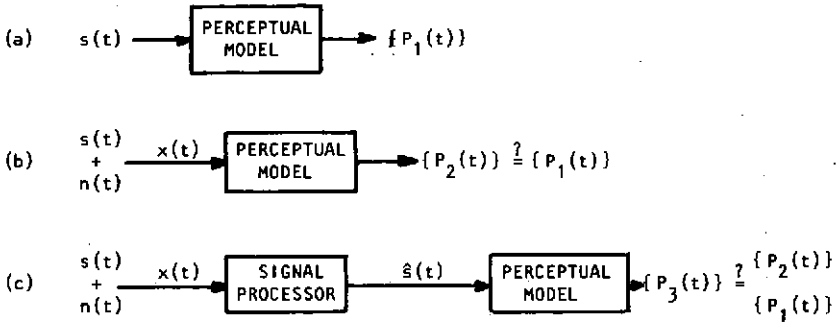

FIG. 1