SPEECH ANALYSIS/SYNTHESIS BASED ON MATCHING THE SYNTHESIZED AND THE ORIGINAL REPRESENTATIONS IN THE AUDITORY NERVE LEVEL

ODED GHITZA*

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139

ABSTRACT

Traditional speech analysis/synthesis techniques are designed to produce synthesized speech with a spectrum (or waveform) which is as close as possible to the original. It is suggested, instead, to match the in-synchrony-bands spectrum measures (Ghitza, ICASSP-85, Tampa FL., Vol. 2, p. 505) of the synthetic and the original speech. This concept has been used in conjunction with a sinusoidal representation type of speech analysis/synthesis (McAulay and Quatieri, Lincoln Laboratory Technical Report 693, May 1985). Based on informal listening, the resulting speech is natural (with some tonal artifact) and highly intelligible both in quiet and noisy environments. The same performance is obtained with two overlapping superposed speech waveforms, music waveforms, and speech in musical background. These results demonstrate the adequacy of the in-synchrony-bands measure in selecting the perceptually meaningful frequency regions of the stimulus spectra. Moreover, the inherent dominance property of this measure significantly reduces the number of sinusoidal components needed for synthesis by approximately 70 percent, offering the potential for reduced data-rate.

INTRODUCTION

Traditional speech analysis/synthesis systems are designed to produce synthesized speech with properties similar to the original. Waveform-coders attempts to produce the original waveform. Vocoder schemes are considered to perform well if they reproduce the short term power spectrum. The deviation from the original depends on the quantization used to achieved a desired data rate.

This report suggests a new speech analysis technique which exploits the limitations imposed by the human peripheral auditory mechanism on speech transduction. It is proposed that speech be synthesized by matching the outputs of the synthesized and the original speech representations at the auditory nerve level, that is at the output of the auditory periphery. In an earlier paper (Ghitza, 1985, [7]) a relative spectral intensity measure was suggested motivated by the in-synchrony characteristics of the timing information in the auditory nerve firing patterns. We termed this measure "the in-synchrony-bands spectrum," SBS in brief. A spectral envelope function was created using the SBS estimates directly, to replace the spectrum path in a channel vocoder. Informal listening to utterances of several male and female talkers suggested that considerable speech information was preserved even though the resulting synthesized speech was of relatively poor quality.

The present study was motivated by the assumption that the poor quality is due to the poor SBS match of the vocoded speech with the original. In order to improve the SBS match, we studied the SBS analyzer using the sinusoidal representation analysis/synthesis system suggested by McAulay and Quatieri, 1985, [10], rather than a vocoder. In the analysis, only the frequency estimates of the SBS (which exhibits the dominance property of the auditory periphery) are used to select the necessary frequency components needed for the sinusoidal representation synthesis. The resulting synthesized speech has an SBS representation identical to the original. Furthermore, the synthesized speech is natural (with some tonal artifact) and highly intelligible for various kinds of acoustical stimuli, both in quiet and noisy environments. These results demonstrate the adequacy of the in-synchrony-bands measure in pointing out the perceptually meaningful frequency regions of the stimulus spectra. The results may also suggest that for analysis/synthesis purposes, the SBS profiles are a sufficient display of speech as represented in the auditory nerve level.

The first two sections describe the sinusoidal speech representation and the in-synchrony-bands spectrum analysis. In the next two sections, the integration of the SBS measure into the sinusoidal representation system is described and the resulting performance of the overall system, termed "the synchrony based dominance system," is demonstrated.

THE SINUSOIDAL REPRESENTATION SYSTEM

In our experiments, we used the simplified version of the sinusoidal representation analysis/synthesis system suggested by McAulay and Quatieri, 1985, [10], termed by the authors as the "overlap and add system" ([10], p. 37). In this simplified version, the speech waveform is modeled as the sum of sine waves; the sine wave parameters are kept constant over the frame. If $s_i(n)$ represents the speech waveform samples in the i'th frame, then

$$s_i(n) = \sum_k A_{ki} \sin(2\pi f_{ki} nT + \theta_{ki}) \qquad (1)$$

where A_{ki} and f_{ki} are the amplitude and frequency of the k'th component in the i'th frame and θ_{ki} is the phase of the k'th frequency component referenced to the center of the i'th frame. In order to determine those parameters, a 512-point FFT is computed every frame and a set of sine wave frequencies is generated by applying simple peak picking to the interpolated magnitude function. The amplitudes and the phases are measured from the interpolated FFT at the location of the

* Currently with the Department of Acoustic Research, AT&T Bell Laboratories, Murray Hill, NJ.

ICASSP 86, TOKYO

peaks. In the synthesizer, the summation (1) is used to produce the frame output signal on which a triangular window is applied. Since the analysis is performed every 10 milliseconds on a 20 milliseconds window, the current and the preceding 20 milliseconds time-weighted frame output signals are overlapped by 10 milliseconds delay and added, to obtain the output synthesized speech. There is no voicing decision and the same procedures are applied for the voiced and the unvoiced frames.

In this section we described the basic system we used for the auditory-based modifications suggested in the following sections. As we shall see later, it has a structure which is very conveniently controlled by the SBS profiles.

THE IN-SYNCHRONY-BANDS SPECTRUM (SBS)

Motivation

All the information that is processed by the higher auditory system stages must exist in the auditory nerve firing patterns, since the auditory nerve is the only link from the auditory periphery to the brain. For speech analysis purposes it is thus sufficient to retain only properties of the speech signal which determine the auditory nerve firing patterns. Measurements of the firing responses of auditory nerve fibers to speech-like stimuli (Sachs and Young, 1979, [11]; Young and Sachs, 1979, [12]; Delgutte and Kiang, 1984, [2]-[6]) suggest that firing rate is an insufficient carrier of speech information. Some use of temporal characteristics of the firing patterns seems necessary. It is also evident from these measurements that as the stimulus intensity increases, more fibers fire in synchrony with the stimulus periodicity. It is thus possible to consider the width of the region in which all the fibers fire in synchrony with the stimulus periodicity as a measure of the stimulus intensity.

Based on these observations, a speech analyzer is suggested. First, the aspects concerning the design of a theoretical speech analyzer are considered; the actual implementation will be presented in part B of this section. The ideal speech analyzer comprises of two stages. The first stage models the peripheral auditory processing structure up to the level of the auditory nerve. The second stage is an heuristic non-linear relative spectrum intensity measure, operates on the output of the first stage and playing the role of the higher level processing. This measure uses timing-synchrony information in an attempt to exploit the in-synchrony phenomena observed in the neuron firing patterns. Its estimation principles are based on conclusions, yet to be psychophysically verified, derived from the gross characteristics of the cat's auditory nerve firing patterns. Specifically, we adopt the temporal non-place approach suggested by Carlson et al, 1975, [1]. It is assumed that information about the relative intensity of different spectral portions of the signal is in the number of fibers which fire synchronously, regardless of the fiber's characteristic frequency. Furthermore, the phase response of the fibers are assumed to be irrelevant. The basis for this assumption is the findings of Goldstein and Srulovicz, 1977, [8], that the interspike interval statistic is adequate to explain the psychophysics of the perception of the pitch of complex tones. The in-synchrony idea was also applied to unvoiced speech with its energy located at the high portion of the spectral band (up to 4-5 kHz). Although synchrony drops as fiber cf increases (Johnson, 1980, [9]), it is assumed that the amount of synchrony in these fibers is still useful.

Implementation

Because of implementation constraints, the design of the first stage (Fig. 1-a) is based on the general overall behavior of the peripheral auditory system. Fine details of the cochlear filters were ignored. The first stage consists of a 100 channel filter-bank, where the filters are highly overlapped and equally spaced in the logarithmic scale with a 3% frequency step. Their frequency responses are similar to the tuning curves of the auditory nerve fibers. Each filter is identified by its

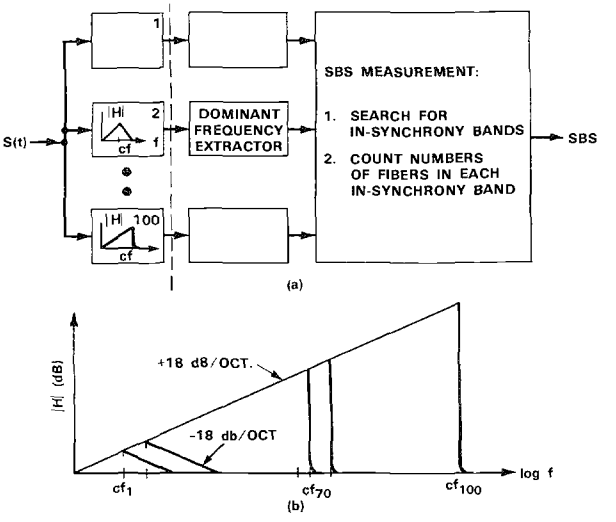

Figure 1

characteristic frequency, cf. The filters with cf up to 1000 Hz have a frequency response with log symmetry around cf, with an +18 dB per octave incline in their low frequency end and a -18 dB per octave roll-off in their high frequency end. The filters with cf above 1000 Hz have a +18 dB per octave incline in their low frequency end but a very sharp roll-off at the high frequency end (Fig. 1-b). In the following text we sometimes refer to such a filter as a "fiber".

The second stage of the proposed speech analyzer is based on the assumptions made in Part A of this section. These assumptions lead to the following definitions:

Definition 1: An in-synchrony-band is a region of L_n successive filters having the same dominant frequency f_n, where the "dominant frequency" is the frequency of the strongest component in the filter's output signal. For a region to be declared as an in-synchrony band, it is necessary that L_n be greater or equal to a threshold, M.

Definition 2: The in-synchrony-bands spectrum (SBS) is a discrete function in frequency, consisting of a set of lines located at frequencies f_n with magnitudes L_n, where f_n and L_n are as in Definition 1.

Time domain implementation of a continuous SBS measurement is clearly computationally complex. Taking into consideration that speech signal characteristics remain steady over approximately 20 milliseconds and can be tracked adequately in a 100 frames per second analysis, we implemented a frame oriented SBS analyzer. The analyzer

operates in the frequency domain and the extraction of the dominant frequency is by picking the frequency of the largest frequency component in the fiber's output power spectrum. The auditory nerve fibers are specified with a simple description in the dB-log frequency scale. In this domain, the fiber filtering can be easily performed by summing the input log spectrum with the log frequency response of the fiber.

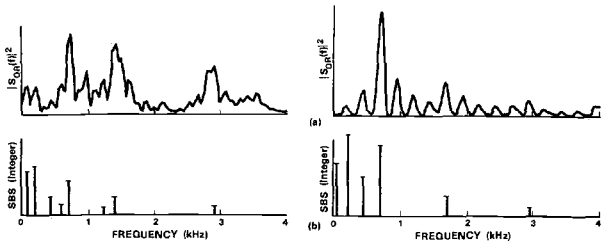

Figure 2

Figure 2-b shows the SBS lines for a sample high resolution speech power spectra (the analysis conditions are described in the Results Section) plotted in Fig. 2-a. The left side plots are samples of male speech while the right side show samples of female speech. Both power spectra are shown after a 6 dB per octave preemphasis. The figures demonstrate the dominance effect which is a basic property of the SBS measure. Stronger frequency components dominate the activity of fibers with higher cf. The magnitude of the SBS lines (obtained by using a highly non-linear operation namely counting the number of in-synchrony fibers) represents the relative importance of each activity region. Figure 2 also shows another characteristic of the SBS. Because of the fiber distribution along the Basilar membrane, most of the frequency components in the first formant region are usually represented in the SBS display. This contributes to the pitch estimation performance of the central processor. The other formants are represented with many fewer SBS lines.

THE SYNCHRONY BASED DOMINANCE SYSTEM

The ability of the SBS measure to serve as an adequate speech description was examined in an earlier study, using listening criteria (Ghitza, 1985, [7]). In that study, the SBS line magnitudes were directly used to create a parallel formant spectral envelope function, where the formant frequencies and their amplitudes were f_n and L_n, respectively. This spectral envelope was used to replace the spectrum path in a channel vocoder. The resulting synthesized speech was of relatively poor quality. The SBS match of the synthesized speech with the original was inadequate, too. However, informal listening to various male and female utterances suggested that considerable speech information was preserved.

Showing that SBS information alone is sufficient should be either by suggesting another transformation of the SBS or by running a synthesis by analysis system that will match the synthesized speech SBS representation to the original. The appropriate transformation of the SBS is not yet been found nor an appropriate control law for the synthesis by analysis system. However, an approximation can be made to modify the SBS by taking into account an *additional,* non SBS, information. It is thus proposed that only the SBS dominance property will be used, to select the meaningful frequency regions in the stimulus spectra. From these regions a complimentary information on the intensity and the phase are to be extracted. Consequently, we are using a modified SBS measure in which the intensity information of each SBS line is replaced by the amplitude of the original speech frequency component at that frequency, computed by a high resolution FFT.

The implementation of the proposed synchrony based dominance analysis/synthesis system is straightforward (Fig. 3). The SBS analyzer is integrated into the analysis side

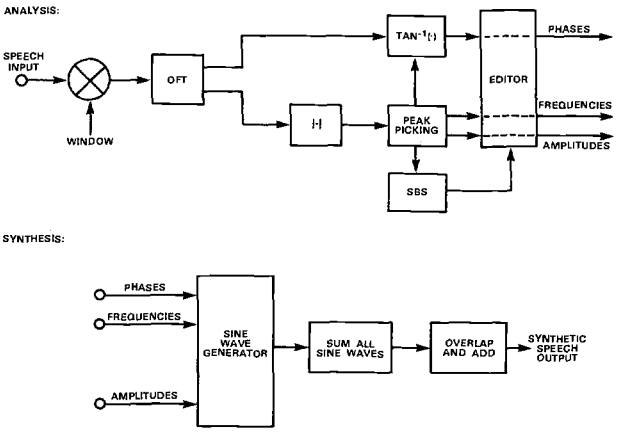

Figure 3

of the sinusoidal representation analysis/synthesis system, to indicate which of the speech frequency components should be transmitted. The transmitted parameters are the SBS line frequencies and the original amplitudes and phases at those frequencies, computed from the speech input spectrum (as in the full sinusoidal representation system). In the synthesizer, only those frequency components are included in the summation of equation (1). The SBS display of the original and the synthesized speech should be identical since the unnecessary frequency components in the original are replaced by zero amplitude frequency components and their is no change in the

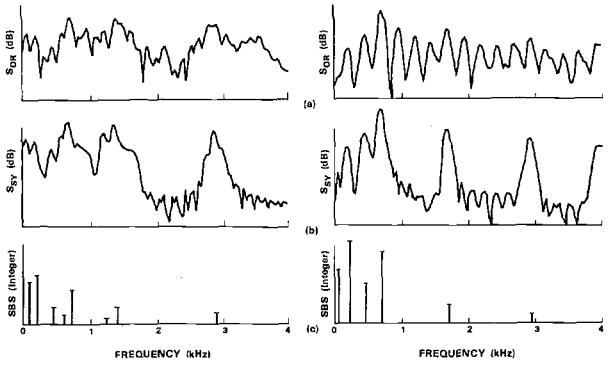

Figure 4

dominance relations. Figure 4-c shows the SBS displays of the original speech spectra (Fig. 4-a) and the synthesized speech spectra (Fig. 4-b) as well. The spectra in Fig. 4-a are the dB versions of the same speech power spectra of Fig. 2-a.

RESULTS

The synchrony based dominance system was applied to a database comprises of a variety of male and female single speaker utterances, two overlapping superposed speech

waveforms, music waveforms and speech in musical background. The speech was low-pass filtered at 5 kHz, preemphasized by a 6 dB per octave preemphasis analog network, digitized at 10 kHz and analyzed at 100 frames per seconds. The analysis was performed using a Hamming window weighted, 20 millisecond frame. The synthesized speech is natural and highly intelligible, but include some tonal artifact. Its SBS display is identical to the original. The distortion is noticeable especially in very low pitched male utterances, where about 80 percent of the frequency components are removed. Two observations have been made. First, compared to the full sinusoidal representation, the average number of the necessary frequency components is about 40 percent for unvoiced frames and about 30 percent for voiced frames. This is due to the fact that in the unvoiced case there are no clear dominance regions since the spectral structure is inharmonic. Secondly, there is about 10 percent more saving for a male talker compared to a female talker. This is due to the lower male pitch which causes a larger density of spectral frequency components to be removed. These results hold for all kinds of input stimuli, including the two overlapping speech waveforms and speech in musical background.

The performance of the system is not seriously affected by the presence of noise, even as high as a 0 dB peak vowel to average noise ratio. In fact, there may be some noise reduction based on the dominance effect. Obviously, each frequency component in the speech frame spectrum is filtered by the narrow band 512-point FFT filter and the high energy frequency components are measured with a better accuracy. Since the low energy frequency components are dropped by the dominance effect and the synthesizer uses only the high energy frequency components, the overall signal to noise ratio may be improved.

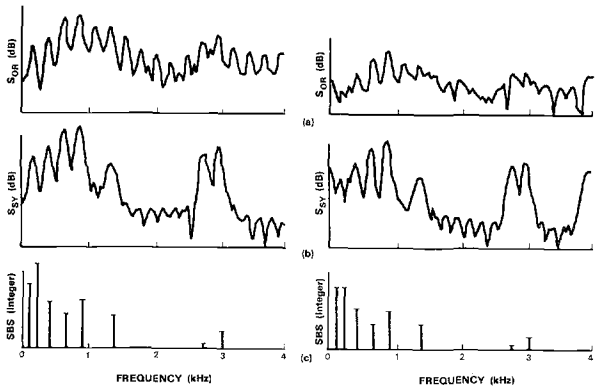

Figure 5

Figure 5 shows the original and the synthesized power spectra and their SBS (Fig. 5-a, 5-b and 5-c, respectively) for a female speech frame in quiet (left plots) and noisy (with a 3 dB peak vowel to average noise ratio) environments. Note that the SBS display is hardly affected by the noise.

In order to get a crude estimate of the upper bound for the number of lines needed per frame, the maximum number of sine wave components to be superposed by the synthesizer was set to 10. Thus, in each frame, the ten SBS components with the largest FFT amplitudes were taken into account. The resulting synthesized speech was perceptually indistinguishable from the unconstrained synchrony based dominance synthesized speech. This result holds for all the stimuli in the database that we described, both in clear and noisy environments.

CONCLUSIONS

Motivated by the in-synchrony characteristics of the timing information in the auditory nerve firing patterns, we used the in-synchrony-bands (SBS) measure to define the perceptually meaningful frequency regions of the stimulus spectra. It seems that for analysis/synthesis purposes, the SBS profiles are a sufficient display of speech as it represented in the auditory nerve level. For all kinds of acoustical stimuli in our database, including single male and female speakers in clear and in noise, two overlapped speech waveforms, music waveforms and speech in musical background, the very few frequency components indicated by the SBS lines are sufficient for an adequate representation of all the speech features. More than half of the SBS lines used are in the region of the first formant, representing the intonation (pitch) information. The other formants are represented with many fewer lines. The SBS representation of voiceless sounds, however, seems to be incomplete, since it introduces tonal artifact to the synthesized speech.

Ten frequency components, at most, are sufficient to obtain natural (with some tonal artifact) and highly intelligible synthesized speech. This suggests a potential for an efficient speech coding system based on this technique.

ACKNOWLEDGEMENT

I wish to thank B. Gold, R. J. McAulay, W. M. Rabinowitz, J. Tierney and T. F. Quatieri for stimulating discussions throughout this work.

REFERENCES

[1] Carlson, R., Fant, G. and Granstrom, B. (1975). "Two formant models, pitch and vowel perception," in: Fant, G. and Tatham, M. A. A. (Eds.), "Auditory analysis and perception of speech," Academic Press, London, pp. 55-82.

[2-6] Delgutte, B. and Kiang, N. Y. S. (1984). "Speech coding in auditory nerve: I., III., IV., V.," J. Acoust. Soc. Am. 75 (3), March, pp. 866-918.

[7] Ghitza, O. (1985). "A measure of in-synchrony regions in the auditory nerve firing patterns as a basis for speech vocoding," ICASSP-85, Tampa, Florida, Vol. 2, March, p. 505.

[8] Goldstein, J. L. and Srulovicz, P. (1977). "Auditory-nerve spike intervals as an adequate basis for aural spectrum analysis," In Evans, E. F. and Wilson, J. P. (Editors) Psychophysics and Physiology of Hearing. Academic-press, London, p. 337.

[9] Johnson, D. H. (1980). "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," J. Acous. Soc. Am., 68 (4), Oct., p. 1115.

[10] McAulay, R. J. and Quatieri, T. F. (1985). "Speech analysis/synthesis based on a sinusoidal representation," Lincoln Laboratory Technical Report 693, May.

[11] Sachs, M. B. and Young, E. D. (1979). "Encoding of steady state vowels in the auditory nerve: representation in terms of discharge rate," J. Acous. Soc. Am., 66(2), Aug., p. 470.

[12] Young, E. D. and Sachs, M. B. (1979). "Representation of steady state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers," J. Acous. Soc. Am., 66 (5), Nov., p. 1381.