SPEECH ENHANCEMENT FOR BANDLIMITED SPEECH

David A. Heide *George S. Kang*
Naval Research Lab Code 5555
Washington DC 20375
heide *or* kang@itd.nrl.navy.mil

ABSTRACT

Throughout the history of telecommunication, speech has rarely been transmitted with its full analog bandwidth (0 to 8 kHz or more) due to limitations in channel bandwidth. This impaired legacy continues with tactical voice communication. The passband of a voice terminal is typically 0 to 4 kHz. Hence, high-frequency speech components (4 to 8 kHz) are removed prior to transmission. As a result, speech intelligibility suffers, particularly for low-data-rate vocoders. In this paper, we describe our speech-processing technique, which permits some of the upperband speech components to be translated into the passband of the vocoder. According to our test results, speech intelligibility is improved by as much as three to four points even for the recently developed and excellent Department of Defense-standard Mixed Excitation Linear Predictor (MELP) 2.4 kb/s vocoder. Note that speech intelligibility is improved without expanding the transmission bandwidth or compromising interoperability with others.

INTRODUCTION

In analog voice transmission, speech bandwidth is limited by the channel bandwidth. In switched public telephone circuits or high-frequency radio channels, speech bandwidth is typically less than 4 kHz. With digital speech transmission, speech bandwidth still remains less than 4 kHz because the wider the speech bandwidth, the higher the data rate required for transmission. The 4 kHz speech bandwidth is standard in virtually all digital voice terminals used by the government and industry.

A speech bandwidth of 4 kHz is acceptable for the vowels spoken by a majority of speakers. The same cannot be said for consonants, particularly fricatives (/s/, /sh/, /ch/, etc.), because their spectra extends above 4 kHz (see Fig. 1). Good reproduction of consonants is vital to a proper understanding of speech because they provide important cues for sentence segmentation.

To improve narrowband speech, we developed a technique to shift the fricative spectrum down below 4 kHz so that some of the fricative sounds will be heard over digital voice terminals. Our earlier effort to spread fricative spectra was by making use of the aliasing effect (i.e., spectral folding around 4 kHz) [1]. This technique is not effective if there is no fricative spectrum in the 4 to 5 kHz region. Thus, we developed a new and more effective technique, which is capable of spreading fricative spectrum even if it is present only above 5 kHz.

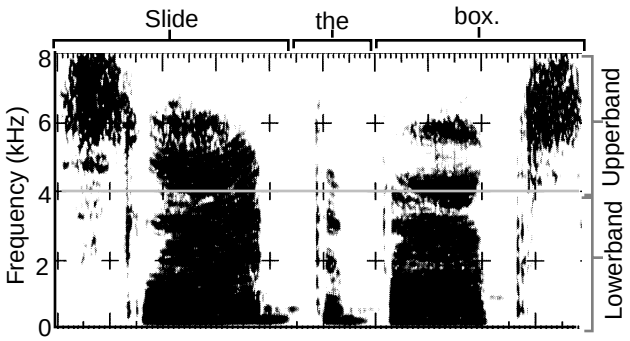


Fig. 1 — Original speech spectrum. Note that the fricative /s/ in /slide/ and /box/ is almost completely above 4 kHz.

FRICATIVE SPREADING TECHNIQUE

Our goal is to spread speech spectra only when they lie mainly above 4 kHz (see Fig. 2) and, otherwise do nothing. In this way, we do not disturb vowels and other low-frequency consonants (such as /p/, /b/, /d/, /th/, etc.). To implement this approach, we decompose the speech spectrum from 0 to 8 kHz into eight equal subbands, a 1 kHz bandwidth for each subband.

Eight-Band Speech Decomposition

We used quadrature mirror filters (QMFs) to achieve speech decomposition [2]. The multiband decomposition by the QMF technique is a repetition of a two-band decomposition in which the input signal is split into a lowerband and upperband. This two-band decomposition requires a low-pass filter $H_1(z)$ and a high-pass filter $H_2(z)$. To allow a perfect recombination, the low-pass and the high-pass filters must be complementary. In addition, the

cutoff characteristic of these filters must be sharp enough to reduce the spectral spill into the neighboring subband. For these reasons, we used a 32-tap QMF [3].

Each single stage decomposition only results in two bands of speech with 4 kHz resolution (0 to 4 kHz and 4 to 8 kHz). To achieve 1 kHz resolution as shown in Fig. 2, we need three stages of decomposition (i.e., an eight-band decomposition), each stage a repetition of the single stage decomposition in which low and high-pass filtering is followed by downsampling. Figure 3 shows a block diagram of eight-band decomposition (i.e., three-stage decomposition).

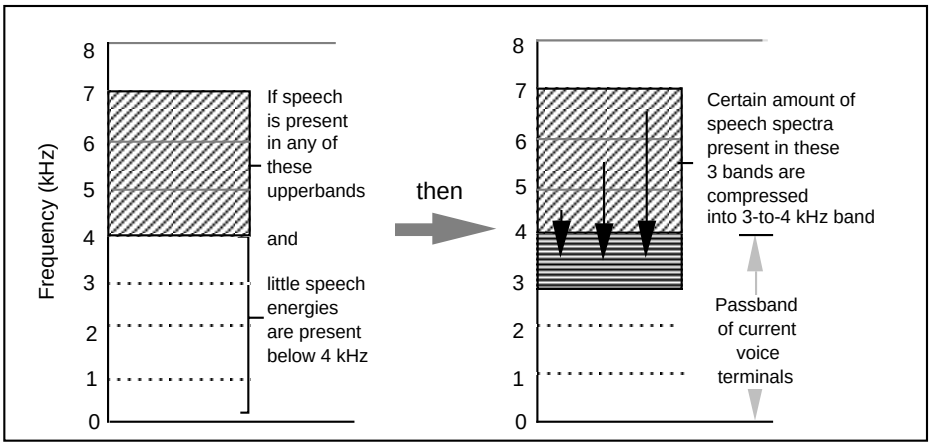


Fig. 2 — Basic concept of spreading of high-frequency spectra into the passband. This spectral translation process is affected only when speech contains primarily high-frequency components (i.e., fricative sounds /s/, /sh/, /ch/, etc.)

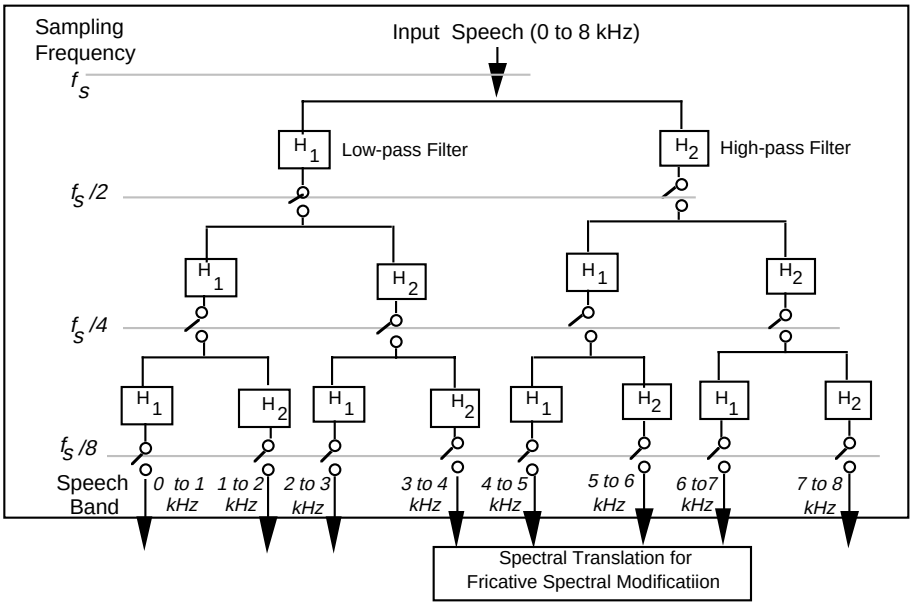


Fig. 3 — Block diagram for splitting speech into eight equal subbands. As noted, an eight-band decomposition is a three-fold repetition of the two-band decomposition.

Once the speech waveform is filtered into eight equal subbands (0 to 1, 1 to 2, 2 to 3, ..., 7 to 8 kHz bands), the remaining process is to spread the fricative spectrum below 4 kHz. The design of such a process depends on the preference of the listener. The approach, however, must be that associated with the enhancement of speech

intelligibility. Thus, we continue to test for speech intelligibility while experimenting with various spectral translation rules. Based on these considerations, this is our approach for spreading fricatives below 4 kHz.

Spectral Translation Rules

1. Compute the speech rms value in each subband: $p(1)$, $p(2)$, $p(3)$, ... , $p(8)$.

2. Compute the spectral centroid (m, m = 1,2,3, ... 8) based on $p(1)$ through $p(8)$. According to our observations, m is around 2 kHz if speech is voiced (vowels), and m is 3 kHz or more if speech is unvoiced (consonants) (see Fig. 4).

3. Spectral translation rules from the upperband to lowerband are:

a. If m is less than or equal to 2 kHz, no spectral translation is performed.
b. If m = 3 kHz, the 4 to 5 kHz spectrum is added into the 3 to 4 kHz band.
c. If m = 4 kHz, the 4 to 6 kHz spectrum is added into the 3 to 4 kHz band
d. If m = 5 kHz, the 4 to 7 kHz spectrum is added into the 3 to 4 kHz band.

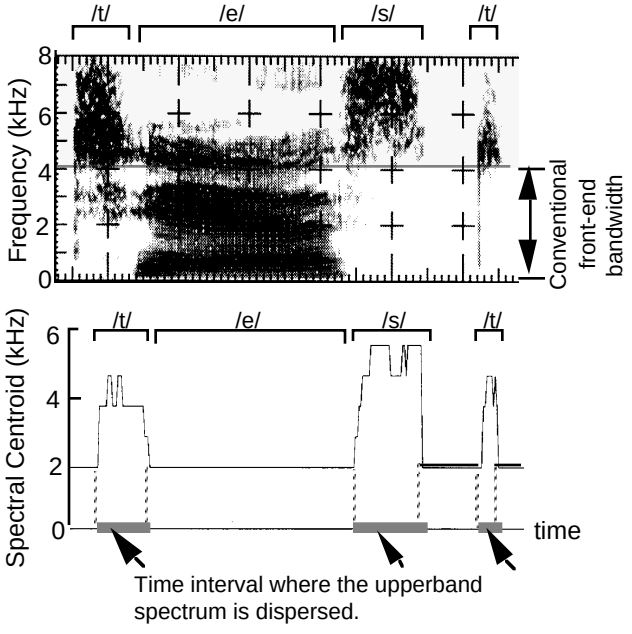
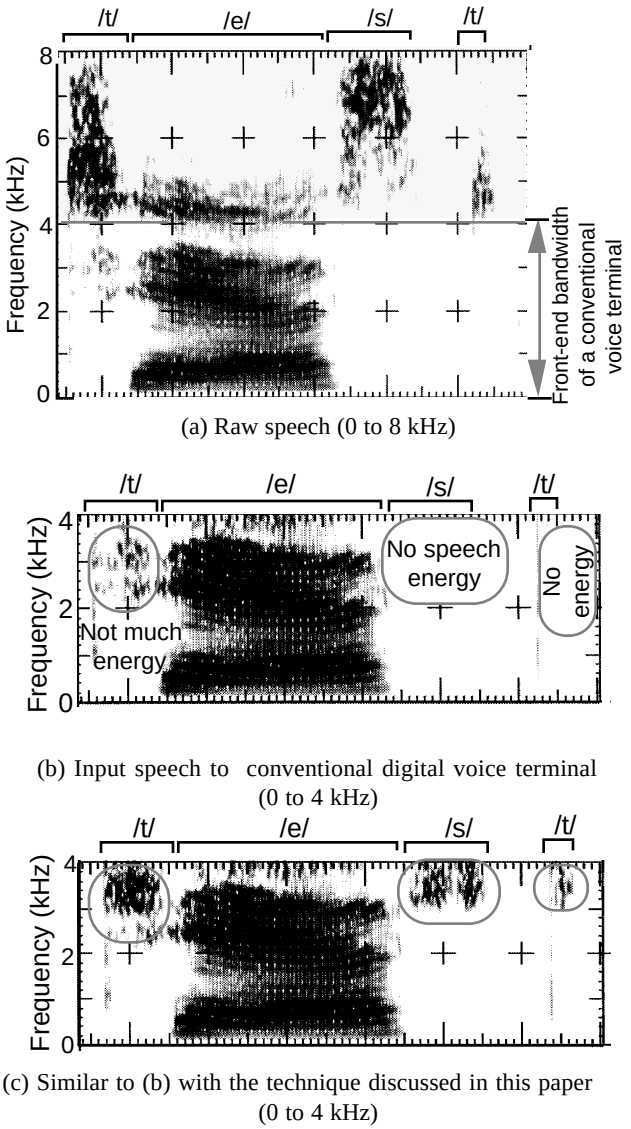


Fig. 4 — Speech spectrum and its spectral centroid.

Figure 5 shows the input and output spectra from this prototype. As noted from Fig. 5(a), a fricative /s/ in /test/ is nearly completely missing below 4 kHz. The fricative enhancement technique downshifts some of the fricative spectrum into the passband so that the listener can perceive the speech correctly, even if the sound quality of the regenerated /s/ is somewhat different from the original /s/. Since people make widely different fricative sounds, slightly altered sounds for fricatives are not that noticeable.



(a) Raw speech (0 to 8 kHz)

(b) Input speech to conventional digital voice terminal (0 to 4 kHz)

(c) Similar to (b) with the technique discussed in this paper (0 to 4 kHz)

Fig. 5 — Input and output speech spectra: (a) is the raw speech spectrum. Note that consonant energies are present mainly above 4 kHz; (b) is the spectrum of the raw speech, which is low-pass filtered at 4 kHz. Since much of the consonant speech energies are eliminated, speech intelligibility is partially lost prior to speech encoding. By the technique discussed in this paper, some of the lost speech energies are recovered as indicated in (c).

INTELLIGIBILITY TEST RESULT

Speech intelligibility before and after fricative spreading is of special interest. Figure 6 shows the result. The intelligibility scores were obtained from the Mixed Excitation Linear Predictor (MELP), which was recently selected as a replacement for LPC-10 and has been in existence since the early 1980s. Test scores indicate that MELP is significantly better than LPC-10. Thus, it is even more difficult to improve MELP, which is already good. It is remarkable that a DRT score for MELP, which is in the "good" range, is enhanced to the "very good" range by the use of the technique presented in this section. Most importantly, this improvement is achieved without expanding the channel bandwidth. In essence, we gained something from nothing, except for some additional computations.

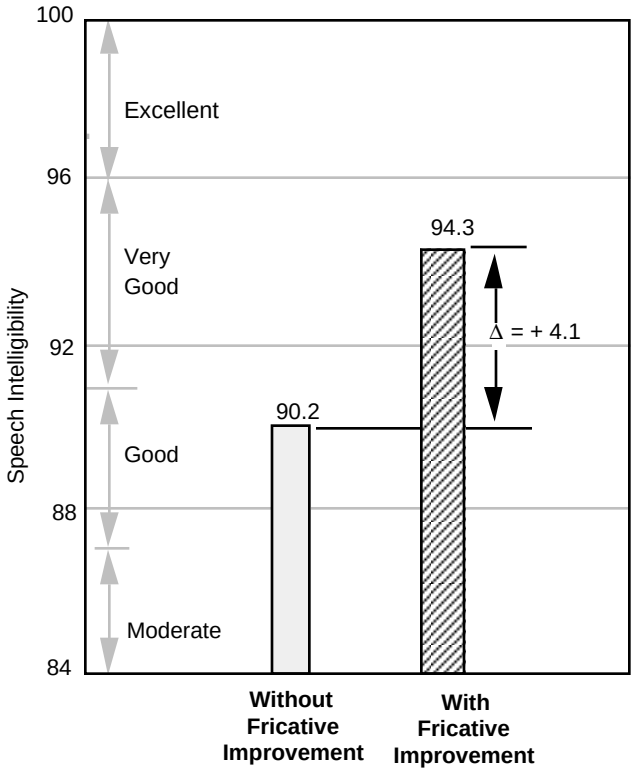


Fig. 6 — Speech intelligibility before and after fricative spreading. The voice encoder is MELP, operating at 2.4 kb/s. The data are obtained from a female speaker talking in a quiet environment.

SUMMARY

In this paper, we described a technique that permits some of the upperband (4 to 8 kHz) speech components to be translated into the 3 to 4 kHz band. As a result, speech intelligibility is improved as much as four points even for the recently developed and excellent 2.4 kb/s MELP. Significantly, speech intelligibility is improved without expanding the passband or compromising interoperability with similar voice terminals that do not incorporate the fricative spreading technique.

ACKNOWLEDGMENTS

The authors thank the NRL Research Advisory Committee for partly supporting these efforts. We also thank CAPT Shupack, Tim McChesney, Phil McCormick of the Navy Secure Voice Office (SPAWAR PMW161) who also supported our research efforts.

REFERENCES

[1] G.S. Kang and S.E. Everett, "Improvement of the Narrowband Linear Predictive Coder: Part 1—Analysis Improvements," NRL Formal Report 8645, Dec. 1982.

[2] D. Estaban and C. Garland, "Application of Quadrature Mirror Filters to Split Band Voice Coding Scheme," *Proc. 1977 IEEE Int. Conf. Acoust. Speech Signal Process*., May 9-11, 1977, Hartford, Conn., pp. 191-195.

[3] M.J. Smith and T.P. Barnwell, "Exact Reconstruction Techniques for Tree-Structured Subband Coders," *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-34**(3), 434-441 (1986).