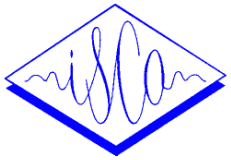


6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Encoded Speech Recognition Accuracy Improvement in adverse Environments by enhancing Formant Spectral bands[1]

S. Kadambe and R. Burns

HRL Laboratories, LLC
3011 Malibu Canyon Road, Malibu CA 90265
E-mail: {skadambe, rburns}@hrl.com

ABSTRACT

Spoken dialogue information retrieval applications are the future trend for mobile users in automobiles, on cellular phones, etc. Due to the limitation of resources in these platforms, it may be advantageous to extract speech features, and compress and transmit them to a central hub where the computation intensive tasks such as speech recognition and speech understanding, etc. can be performed. Generally, the speech recognition accuracy degrades when the decoded speech signal (that is obtained after re-synthesizing the signal from the compressed features) is used. In addition, the background noise that is present in the above mentioned mobile systems will reduce the recognition accuracy. Therefore, in order to improve the recognition accuracy it is essential to extract robust features that can jointly optimize compression and recognition. In this paper, we describe a technique that improves the recognition accuracy of noisy encoded speech signals by performing spectral correction and spectral formant band enhancement before synthesizing the speech signal from the compressed features. We have conducted experiments on 1831 telephone speech utterances from 1831 speakers. We added (a) the in-vehicle noise recorded from a Volvo car moving on an asphalt road at 134 kmph, (b) the factory noise recorded in a factory and (c) the speech (babble) noise recorded in a cafeteria to these utterances at various signal-to-noise ratios (SNR). Our experimental results indicate recognition accuracy improvement up to 10% at 0 dB SNR.

1 INTRODUCTION

Spoken dialogue based information retrieval applications are the future trend for mobile users in automobiles, on cellular phones, etc. Due to the limitation of resources in these mobile systems, it is advantageous to extract speech features, and compress (code) and transmit them to a central hub where the computation intensive tasks such as speech recognition and speech understanding, etc. can be performed. Generally, the speech recognition accuracy degrades when a decoded speech signal is used. This is true even in the case of uncorrupted coded and decoded speech signals. In addition, the background noise that is present in the above mentioned mobile systems will reduce the recognition accuracy. To mitigate the problem of noise in mobile systems several techniques have been developed. In [1], the authors assume that the likely acoustical environment is known *a priori* and have developed Adaptive Vector Quantization technique to adapt to the new test environment. In [2], the authors first apply noise reduction techniques in the homomorphic domain and then compute speech features. They use radial basis neural network as the final classifier. Even though the results of these two studies are significant, they did not consider the effect of encoded speech. In [3], the authors considered the encoded speech features and showed that by concatenating feature vectors derived from the encoded GSM parameters they can improve the speech recognition accuracy by 4.4% as compared to the speech recognition accuracy of the original waveforms using Mel-Frequency Cepstral Coefficients (MFCC) features. This improvement is in the case of noisy signal with 18 dB SNR.

In this paper, however, we describe a technique that improves the continuous speech recognition accuracy of noisy encoded speech at low SNRs. While performing the recognition, the signal is recovered from the encoded speech features. The encoded speech features get corrupted by background noise and transmission channel. Therefore, before recovering the signal, the corrupted encoded speech features are corrected by applying the proposed spectral correction and formant spectral band enhancement techniques. For encoding of speech features, the LPC model based Sony's encoder-decoder [4] is used. A brief description of this encoder-decoder is provided in the next section. The proposed spectral correction and formant spectral band enhancement techniques are described in section 3. Simulation details and recognition experimental results are provided in section 4. We conclude and discuss the future research directions in section 5.

2 A BRIEF DESCRIPTION OF SONY's ENCODER/DECODER

Sony's encoder-decoder uses an all-pole model for the extraction of speech features. Such an all-pole model based technique is popularly referred to as LPC analysis in speech literature. A given speech signal is analyzed using such a model. The features (formant frequencies and their bandwidth) of this speech signal are extracted by estimating the roots of an optimum model that is obtained after analysis. The roots of the model provide the information about the pole locations and they correspond to formant frequencies. In the frequency or spectral domain, these poles correspond to peaks. Hence, the frequencies associated with peak locations correspond to formant frequencies, and the widths of the peaks correspond to formant bandwidths.

Sony's decoder-encoder first classifies a given frame of a speech signal into voiced or unvoiced. If the signal is voiced, vector quantization of harmonic spectral envelope of LPC residuals with a weighted distortion measure is used whereas if the signal in unvoiced, vector excitation with stochastic codebooks is

used. While decoding, i.e., re-synthesizing the speech signal from the coded parameters, sinusoidal excitation source is used for voiced signals and a codebook lookup table is used for unvoiced speech signals.

3 ENHANCEMENT OF CORRUPTED ENCODED SPEECH FEATURES

As mentioned before, both background noise and transmission channel corrupts encoded speech features. These are corrected by applying the proposed spectral correction and formant spectral band enhancement techniques. These two techniques are described in section 3.1 and 3.2, respectively.

3.1 Spectral correction

One of the effects of noise (both channel effect and additive noise) in the LPC model based speech analysis is that the locations of the poles get perturbed as compared to the clean signal. To reduce the noise effect and hence to improve the speech recognition accuracy, the first step that is used in this paper is to apply the spectral correction technique. This consists of adjusting the locations of poles i.e., moving them inside the unit circle. Such a movement of poles can cause instability. To over come this, a zero is placed in such a location that it offsets the instability caused by the movement of a pole location.

3.2 Formant spectral band enhancement

From the spectral corrected LPC spectrum, bandwidths of formant frequencies can be estimated as shown pictorially in Figure 1. A finite impulse response (FIR) band pass filter corresponding to each format is designed using a standard filter design technique. The center frequency of each band pass filter corresponds to each formant frequency (f_i or $\omega_i = 2\ \pi\ f_i$)and bandwidth corresponds to bandwidth of each formant frequency ($\Delta\omega_{f_i}$). The LPC spectrum is filtered using these set of band pass filters. The output log power spectrum is compared with the log power spectrum of the uncorrupted encoded speech features using the spectral distance measure (SD)

$$SD_{f_i} = \frac{1}{\Delta\omega_{f_i}} \sum_k \sqrt{\left[\log\left(S_{f_i}(\omega_k)\right) - \log\left(\hat{S}_{f_i}(\omega_k)\right)\right]^2}$$ to

verify the spectral enhancement in each of the formant spectral band. Here S & $\hat{S}$ correspond to the power spectrum of uncorrupted and corrected encoded speech features within the bandwidth $\Delta\omega_{f_i}$ of formant f_i and the summation is over the frequencies that fall within the bandwidth of f_i. If the mean SD is within 1 dB, it is assumed that the required enhancement is obtained.

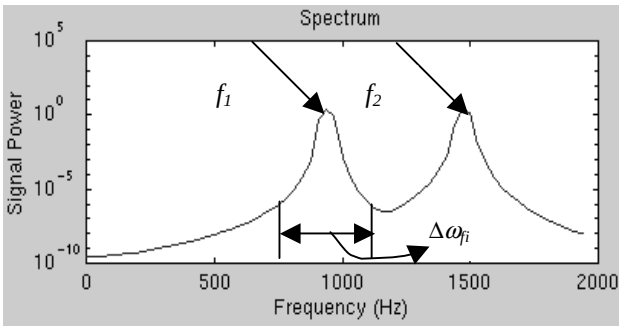
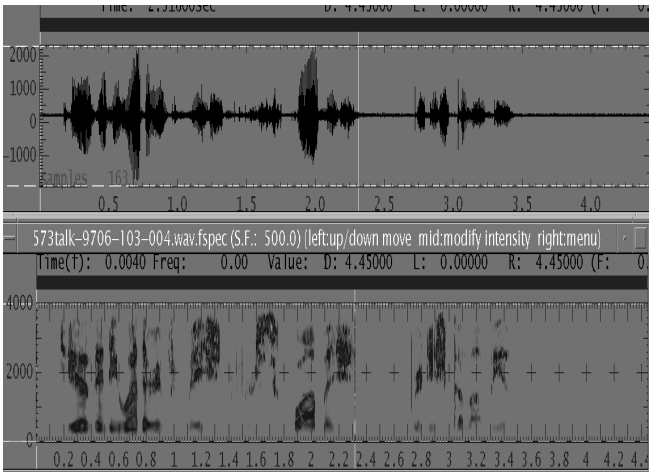


Figure 1: Pictorial description of formant frequencies and formant band width

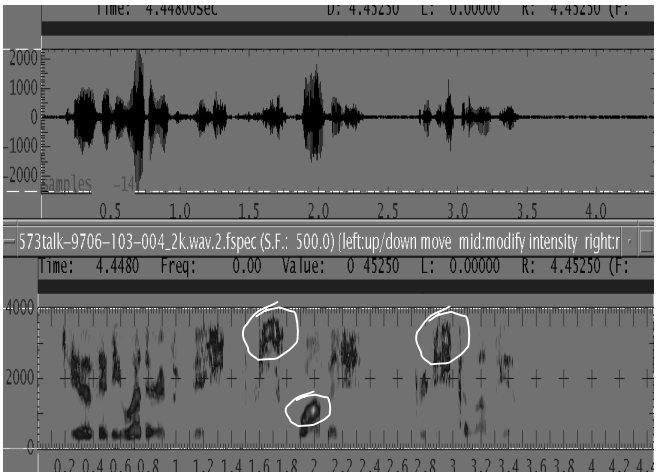
4 SIMULATIONS

4.1 Verification of formant enhancement

First, to verify enhancement of spectral formant bands, a clean speech signal was considered. Speech features of this signal were extracted and encoded using Sony's encoder. This data was used in obtaining the LPC spectrum of uncorrupted speech features. The encoded speech features were then corrupted by perturbing certain bits. Spectral correction and formant band enhancement techniques described in the previous section were applied. The signal was then synthesized using the corrected LPC spectrum. For the synthesis, Sony's decoder was used. In Figure 2, a clean speech signal, and its spectrogram (Figure 2 (a)), and, re-synthesized signal with correction and its spectrogram (Figure 2 (b)) are plotted, respectively. From this figure, it can be seen that the intended formants (marked regions in the spectrogram of Figure 2 (b)) have been enhanced as compared to the spectrogram of the original un-encoded speech signal.



(a)

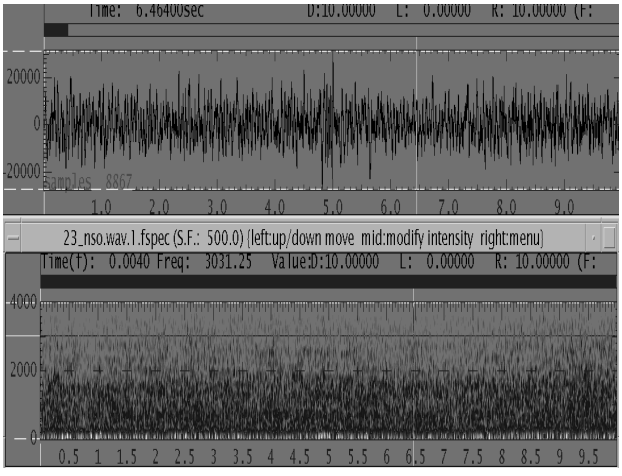


(b)

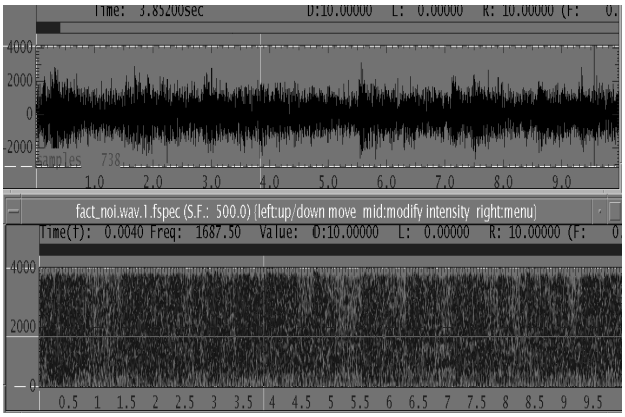
Figure 2: (a) Clean speech signal and it spectrogram, and (b) re-synthesized speech signal after spectral correction and enhancement and its spectrogram, respectively. Some of the enhanced formant bands are marked in the spectrogram of figure (b)

4.2 Verification of speech recognition accuracy improvement

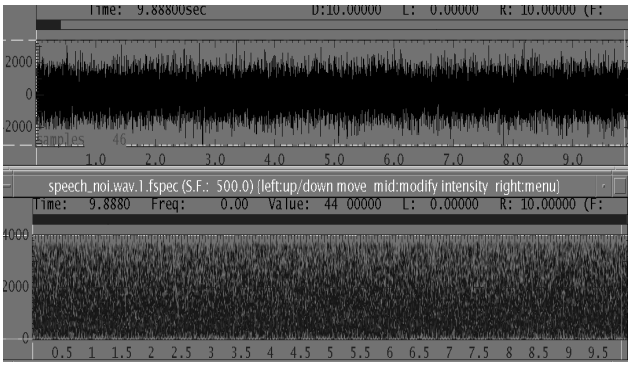
Experiments were conducted to verify whether the spectral correction and formant band enhancement techniques would improve the recognition accuracy of encoded speech signals if they were corrupted by different additive noises. Note that for brevity, from now on we refer to the process of spectral correction and formant band enhancement as "enhancement". For this set of experiments, three different noises – in-vehicle noise recorded from a Volvo car moving at the speed of 134 kmph on an asphalt road, factory noise and speech (babble) noise were considered. These three noise signals and their respective spectrograms are plotted in Figure 3 (a) (b) and (c), respectively.



(a)



(b)



(c)

Figure 3: (a) In-vehicle noise and its spectrogram, (b) factory noise and its spectrogram, and (c) speech (babble) noise and its spectrogram, respectively.

From these figures, it can be seen that the most difficult case is the speech noise since it exhibits spectral energy at all frequencies. The next difficult case is the factory noise since it exhibits periodic nature similar to speech and it has spectral energy at the frequencies that are close to unvoiced speech signals. The least difficult case out of three types of noise is in-vehicle noise. The in-vehicle noise exhibits spectral energy at low frequencies that may overlap with pitch frequencies of voiced speech signals. The experimental details and results of recognition accuracy using these three types of noises are provided below.

For recognition experiments, 1831 utterances from 1831 speakers were considered. This data was collected over the telephone channel by Spoken language laboratory of MIT as part of their Jupiter system. First, the recognition accuracy for 1831 clean speech utterances was obtained by using a continuous speech recognizer [5]. This recognizer is segment based and uses (a) di-phone acoustic models that are trained using MFCC based speech features and mixture Gaussian distribution, (b) a pronunciation dictionary that is generated using specific phonological rules and (c) a bi-gram language model. These 1831 speech utterances were then encoded and decoded after enhancement using Sony's encoder-decoder described above. Compression rates of 2 kbps and 4 kbps were used for encoding. Table 1 provides the recognition accuracy for all these three cases. From this table it can be seen that for no

noise case the enhanced decoded signals did not reduce the recognition accuracy significantly.

	No encoding-decoding	Encoding-decoding with enhancement at 4 kbps	Encoding-decoding with enhancement at 2 kbps
Word recognition accuracy	91.9 %	88.1 %	85.7 %

Table 1: The recognition accuracy of clean speech signals with and without encoding-decoding and enhancement

Next, in-vehicle, factory and speech noises were added to all these 1831 utterances, respectively, such that the SNR of resulting noisy speech signals is equal to 0 dB. The recognition experiment was conducted on all three different noisy speech files (1831 utterances) with and without encoding-decoding and enhancement. The recognition accuracy for each noise case is provided in Table 2. From this table it can be seen that the recognition accuracy degrades with severity of noise (least recognition accuracy for speech noise). However, encoding-decoding with enhancement improved the recognition accuracy in all three cases. The improvement is the highest (9.7 %) for the difficult case of speech noise and lowest (5.2 %) for comparatively easy case of car noise. This implies that if noise is more like speech then the recognizer can benefit more from the enhancement technique of this paper.

Noise type	0 dB SNR, no encoding-decoding	0 dB SNR with encoding-decoding and enhancement at 4 kbps
In-vehicle noise	44.7 %	49.9 %
Factory noise	25.8 %	32.3 %
Speech noise	22.9 %	32.6 %

Table 2: The recognition accuracy of noisy speech signals with and without encoding-decoding and enhancement

To verify it is indeed the spectral enhancement of formant bands (that can be estimated from the LPC spectrum that the Sony encoder-decoder computes) help in improving the recognition accuracy of noisy speech signals, a coder based on multiple excitation cepstral coefficient (MCELP) was used. This coder does not treat voiced and unvoiced sounds differently and does not compute the LPC spectrum that can be used for spectral enhancement of formant bands. This is the more commonly used compression technique. In Table 3, the recognition results of clean and two types (in-vehicle and speech noise) of noisy speech signals with and without MCELP encoding-decoding are provided. From this table, it can be seen that the decoded speech from MCELP coder-decoder degrades the recognition accuracy in all cases. This further substantiates that the enhancement of formant spectral bands described in this paper helps in reducing the effect of especially speech like noise on recognition accuracy.

Noise type	0 dB SNR, no encoding-decoding	0 dB SNR, MCELP based encoding-decoding at 4 kbps
In-vehicle noise	44.7 %	38.1 %
Speech noise	22.9 %	19.8 %

Table 3: The recognition accuracy of noisy speech signals with and without MCELP encoding-decoding

Note that the speech recognizer used in this study is not trained for noisy speech signals. The recognition accuracies would be much higher if the training & test environments matched. Note also that the emphasis of this paper is how to improve the speech recognition accuracy in mismatched case and not to validate any given speech recognizer.

5 CONCLUSIONS

In this paper, a technique based on spectral correction and formant spectral band enhancement to improve the recognition accuracy of encoded speech is described. Three types of noises – in-vehicle, factory and speech (babble) are considered. From the spectrogram of these noises, it can be seen that the speech noise has the most overlap of spectral energy with speech signals. The speech recognition accuracy experiments were conducted using a continuous speech recognizer. From the recognition results it can be seen that by applying the proposed enhancement technique, the recognition accuracy of encoded noisy speech signals can be improved. The improvement is the highest for the most severe case i.e., speech noise. Even though the recognition accuracy is improved by 10 % for speech noise (0 dB SNR), this improvement is still low for any practical application of speech recognizers in highly noisy environment. Therefore, future research warrants further improvement of recognition accuracy.

ACKNOWLEDGEMENT

The authors would like to thank Spoken Language systems laboratory, MIT, Cambridge, MA for providing the test data that is used in this study.

6 REFERENCES

[1] M. K. Sonmez, R. Rajasekaran and J. S. Baras, "Robust recognition of Cellular telephone by adaptive Vector Quantization," *ICASSP-96*, pp. 503-506.

[2] R. Sankar and N. S. Sethi, "Robust recognition techniques using a radial basis function neural network for mobile applications," *IEEE SOTHEASTCON 97*, pp. 87-91.

[3] J. M. Huerta and R. M. Stern, "Speech recognition from GSM codec parameters," *ICASLP-98*, pp. 626-629.

[4] M. Nishiguchi and J. Matsumoto, "Technical description of Sony IPC's proposals for MPEG-$ audio and speech coding," MPEG-96/0731.

[5] J. Glass, T. Hazen and L. Hetherington, "Real-time telephone based speech recognition in the Jupiter domain," *Proc. Of ICASSP*, Phoenix, AZ, 1999, pp. 61.64.