FEATURE ENHANCEMENT FOR A BITSTREAM-BASED FRONT-END IN WIRELESS SPEECH RECOGNITION

Hong Kook Kim and Richard V. Cox

AT&T Labs-Research, 180 Park Avenue, Florham Park, NJ 07932, USA
E-mail: hkkim@research.att.com, rvc@research.att.com

ABSTRACT

In this paper, we propose a feature enhancement algorithm for wireless speech recognition in adverse acoustic environments. A speech recognition system is realized at the receiver side of a wireless communications system and feature parameters are extracted directly from the bitstream of the speech coder employed in the system. The feature parameters are composed of spectral envelope and coder-specific information. The proposed feature enhancement algorithm incorporates feature parameters obtained from the decoded speech and an enhanced version into the bitstream-based feature parameters. Moreover, the coder-specific parameters are improved by reestimating the codebook gains and residual energy from the enhanced residual signal. HMM-based connected digit recognition experiments show that the proposed feature enhancement algorithm significantly improves recognition accuracy at low SNR without causing poorer performance at high SNR.

1. INTRODUCTION

Wireless communications networks provide more adverse environments for speech recognition than wireline networks. The adverse environments are characterized by low signal-to-noise ratio (SNR) speech due to ambient noise [1] and by packet loss due to channel impairments [2]. The former environment also occurs in wireline networks and results in degrading speech recognition performance [1]. In addition to the packet loss, low-bit-rate speech coders employed in wireless networks result in additional degradation to recognition performance [3]. Moreover, in acoustic noisy environments, a speech coder fails to generate high quality speech. As a remedy to these problems, we can consider introducing robust techniques that are conventionally used in wireline speech recognition [4]. They include speech enhancement, feature normalization/transformation, and model compensation. A typical speech enhancement algorithm tries to subtract the noise spectrum from the noisy speech spectrum by using a spectral subtraction method. Feature normalization techniques such as cepstral mean subtraction (CMS) and a high-pass filtering of cepstral coefficients have been developed to remove the convolutional noise caused by the channel, where the channel includes the transducer effect and telephone channel characteristics [1]. Model compensations or adaptations are used to newly estimate the model parameters with a small amount of data obtained from a new environment. Also, they can be implemented without any additional data. However, model compensation techniques are out of the scope of this work because we are mainly concerned with the design of a robust front-end for wireless speech recognition.

In this work, we assume that a speech recognition system is constructed in the receiver side of a wireless network and the recognition feature parameters are directly obtained from the bitstream generated by a speech coder [5]. Therefore, the robustness of the feature parameters is highly dependent on the accuracy of spectral analysis used in the speech coder. When speech enhancement is applied to the input noisy speech as a preprocessing stage to speech coding, the speech encoder can obtain a more accurate spectral analysis of the noisy speech. And thus, improved performance of the speech recognition system can be achieved [6]. However, speech enhancement may degrade the speech quality in a high SNR environment, and may lower recognition accuracy as well. Moreover, it may not be practical to implement a speech enhancement algorithm at the encoder side of the speech coder. The speech recognition system considered in this work must give reliable performance for speech signals obtained from a broad range of wireless phones. Some phones might have a speech enhancement algorithm, but others may not. This leads us to develop an enhancement scheme that only operates in the receiver side of a wireless communications system.

Following this introduction, we briefly review the bitstream-based front-end for wireless speech recognition in Section 2. In Section 3, we describe the proposed feature enhancement algorithm and discuss its implementation issues. In Section 4, we perform connected digit recognition in additive noisy environments. Finally, we present our conclusions in Section 5.

2. BITSTREAM-BASED FEATURE EXTRACTION

Throughout this work, we choose the IS-136 digital cellular system as a wireless communications system that employs the IS-641 speech coder [7]. The bitstream for a frame is largely divided into two classes. One is 26 bits for line spectrum pairs (LSP) quantization and the other is 122 bits for residual information. We first obtain the decoded LSP's from the 26 bits, where these LSP's represent the spectral envelope of a 30 ms speech segment with a frame rate of 50 Hz. Since the frame rate of the conventional spectral analysis used for speech recognition is usually 100 Hz and we want to simply replace the conventional front-end with the proposed one, we interpolate these LSP's with those of the previous frame and thus make the frame rate 100 Hz. Next, cepstral coefficients of order ten are obtained from the conversion of LSP-to-LPC followed by LPC-to-cepstrum conversion. We obtain the ten liftered cepstral coefficients by applying the bandpass lifter to the cepstral coefficients. For the coder-specific parameters we decode the residual signal from the adaptive codebook lags, the shape or algebraic codebook indices, and the codebook gains. Then, we compute an energy parameter by taking the logarithm of the square-sum of the residual for 10 ms. In addition, the adap-

tive codebook gain and the fixed codebook gain parameters are extracted twice every frame. As a result, the feature vector is 13-dimensional including ten LPC-cepstral coefficients, a normalized logarithmic residual energy, an adaptive codebook gain parameter, and a fixed codebook gain parameter. Finally, we apply CMS to each feature parameter except the logarithmic residual energy, and add the first and second time differences of each parameter, where the first and the second differences are computed by five and three frame windows, respectively. Therefore, the feature vector dimension is 39.

As a reference front-end for comparison, we follow a feature extraction procedure described in the literature [8]. This front-end directly utilizes speech signals recorded over the public switched telephone network (PSTN), and thus we will refer to it as the *conventional wireline front-end*. Speech signals are preemphasized by using a first order differentiator of $1-0.95z^{-1}$. After that, a Hamming window of length 30 ms is applied to each speech segment to obtain a linear prediction polynomial by using the autocorrelation method. The Levinson-Durbin recursion is applied to the autocorrelation coefficients to extract the LPC coefficients of order ten. Finally, the ten LPC coefficients are converted into the twelve cepstral coefficients, and then the bandpass cepstral lifter is applied to the cepstral coefficients. The analysis is repeated once every 10 ms, which results in a frame rate of 100 Hz. A feature vector is 39-dimensional including 12 LPC-cepstral coefficients postprocessed with CMS, a normalized logarithmic energy, and their first and second time differences. The first and the second differences are also computed by five and three frame windows, respectively. A recognition system can be constructed by using the decoded speech signals obtained from the IS-641 speech coder on the receiver side of the IS-136 digital cellular system. The front-end in this system will be the same as the wireline front-end described above, and it will be referred to as the *conventional wireless front-end*.

3. FEATURE ENHANCEMENT

In this section, we propose a feature enhancement algorithm that works in the receiver side of the wireless communications system. The feature enhancement algorithm is realized by utilizing the decoded speech and its enhanced speech, as shown in Fig. 1. The spectral envelope parameters are enhanced in the LSP domain by using the algorithm described in Section 3.1. Also, the coder-specific parameters are also updated by one of two methods: a direct assignment method and a reestimation method, which are explained in Section 3.2. Finally, we propose a SNR-based enhancement method in Section 3.3 in order to avoid degrading the recognition performance under high SNR conditions. In other words, the proposed feature enhancement algorithm is selectively applied to speech signals depending on the estimated SNR. To begin with, we define notation used in describing the feature enhancement algorithm.

3.1. Spectral Parameters

As shown in Fig. 1, we first define four p-dimensional LSP column vectors at the l-th frame.

- $\omega_{n,l}$: an LSP vector derived from the bitstream of the speech coder.

- $\omega_{n+q,l}$: an LSP vector obtained from the decoded speech signal, which is also distorted by both a background noise and quantization error caused by the speech coder.

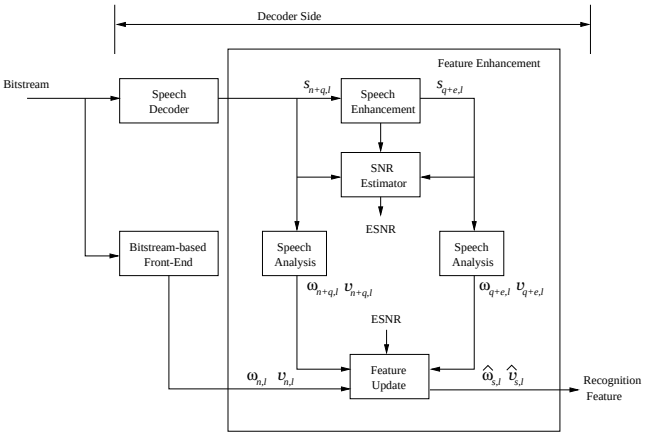

Fig. 1. Block diagram of the proposed feature enhancement algorithm realized at the decoder side of the speech coder.

- $\omega_{q+e,l}$: an LSP vector obtained from the enhanced speech signal[1] at the decoder side of the speech coder. In other words, it is an enhanced version of $\omega_{n+q,l}$.

- $\hat{\omega}_{s,l}$: an enhanced LSP vector after applying the proposed feature enhancement algorithm.

The objective of the proposed feature enhancement algorithm is to find an enhanced LSP vector, $\hat{\omega}_{s,l}$, such that

$$D(\omega_{s,l}, \hat{\omega}_{s,l}) \leq D(\omega_{s,l}, \omega_{n,l}) \tag{1}$$

where the distance $D(\cdot,\cdot)$ is defined by $D(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})$, and $\omega_{s,l}$ means an LSP vector obtained from the actual clean speech signal. However, it is hard to find a closed form solution satisfying the condition of (1). Thus, we propose a feature enhancement algorithm having the form of

$$\hat{\omega}_{s,l} = \omega_{n,l} + f(\omega_{n+q,l}, \omega_{n,l}) + g(\omega_{q+e,l}, \omega_{n+q,l}) \tag{2}$$

where $f(\omega_{n+q,l}, \omega_{n,l})$ and $g(\omega_{q+e,l}, \omega_{n+q,l})$ are the compensation factors for the speech coding effect and background noise, respectively. By introducing a gradient descent algorithm for the compensation, the functions $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ can be represented as

$$f(\omega_{n+q,l}, \omega_{n,l}) = -\frac{1}{2}\mu_f \nabla_{\omega_{n,l}} D(\omega_{n+q,l}, \omega_{n,l}) \tag{3}$$

$$g(\omega_{q+e,l}, \omega_{n+q,l}) = -\frac{1}{2}\mu_g \nabla_{\omega_{n,l}} D(\omega_{q+e,l}, \omega_{n+q,l}) \tag{4}$$

where ∇ denotes a gradient operator, and μ_f and μ_g are smoothing coefficients. The gradient of $D(\omega_{n+q,l}, \omega_{n,l})$ with respect to $\omega_{n,l}$ can be assumed to zero because the speech decoder can be considered as a vector quantizer and this quantizer should satisfy the necessary condition for local minima [10]. Therefore, $\partial\omega_{n+q,l}/\partial\omega_{n,l} = \mathbf{1}$, where $\mathbf{1}$ is a column vector whose elements are all 1. This gives the following equation of

$$\omega_{n+q,l} = \omega_{n,l} + \mathbf{c}_l \tag{5}$$

[1]The speech enhancement algorithm used in this work is based on minimum mean-square error log-spectral amplitude estimation, and it has been applied to some standard speech coders [9].

where $\mathbf{c}_l$ is a constant vector for $\omega_{n,l}$. By substituting (5) into (4), we obtain

$$\frac{\partial D(\omega_{q+e,l}, \omega_{n+q,l})}{\partial \omega_{n,l}} = -2(\omega_{q+e,l} - \omega_{n+q,l}) \quad (6)$$

where we also assume that $\omega_{q+e,l}$ is independent on $\omega_{n,l}$ because the speech enhancement algorithm applied to the decoded speech sufficiently removes the noise component in the noisy speech. By substituting (6) and (4) into (2), we obtain the following equation of

$$\hat{\omega}_{s,l} = \omega_{n,l} + \mu_g(\omega_{q+e,l} - \omega_{n+q,l}). \quad (7)$$

We have the following remarks on the smoothing coefficient and the filter stability.

1) *Determination of μ_g:* By substituting (5) into (7), (7) is rewritten by

$$\hat{\omega}_{s,l} = (1 - \mu_g)\omega_{n,l} + \mu_g\omega_{q+e,l} - \mu_g\mathbf{c}_l. \quad (8)$$

From the sufficient condition for stability of the least mean-square algorithm [11], we know that $0 < \mu_g < 2/\mathrm{tr}(\mathbf{I}) = 2/p$, where $\mathrm{tr}(\cdot)$ is the trace of a matrix and p is the dimension of $\omega_{n,l}$. Therefore, we set $\mu_g = 2/p$ for the fastest convergence. In this work, p = 10, and thus μ_g is set to 0.2.

2) *Filter Stability:* We apply the stabilized procedure to the enhanced LSP vector, $\hat{\omega}_{s,l}$, by simply checking the ordering property of LSP [12]. If the enhanced LSP vector is decided as unstable, we replace the vector with the original one. In other words, we set $\hat{\omega}_{s,l} = \omega_{n,l}$. By doing this, we can solve the stability problem of the feature enhancement algorithm.

3.2. Coder-Specific Parameters

From now on, we explain how to enhance the coder-specific parameters such as the residual energy, the adaptive codebook gain parameter, and the fixed codebook gain parameter.

3.2.1. Direct assignment

The additive noise also blurs the voicing information used in the bitstream-based front-end. From the listening tests with noisy and enhanced speech signals, we found that enhancing the noisy speech signal improves the voicing information as well as the spectral information. Therefore, we can simply update the enhanced coder-specific parameter vector as

$$\hat{\mathbf{v}}_{s,l} = \mathbf{v}_{q+e,l} \quad (9)$$

where $\hat{\mathbf{v}}_{s,l}$ and $\mathbf{v}_{q+e,l}$ are the coder-specific parameter vectors corresponding to $\hat{\omega}_{s,l}$ and $\omega_{q+e,l}$, respectively.

3.2.2. Residual enhancement-based parameter reestimation

As shown in Fig. 2, let $r(k)$ and $r'(k)$ be the residual signal and the enhanced residual signal, respectively. Then, the adaptive codebook gain is reestimated as

$$g_p'(i) = \frac{\sum_{k=iN_s}^{(i+1)N_s-1} r'(k-T_i)r'(k)}{\sum_{k=iN_s}^{(i+1)N_s-1} r'(k-T_i)r'(k-T_i)}, i = 0, \cdots, 3 \quad (10)$$

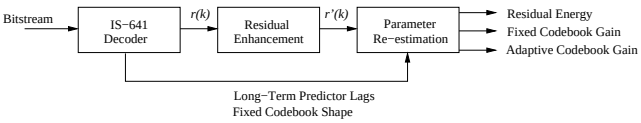

Fig. 2. Procedure of reestimating the coder-specific parameters.

where N_s is the subframe size of 40 and T_i is the long-term predictor lag of the i-th subframe. Also, for $i = 0, \cdots, 3$, the reestimated fixed codebook gain is given by

$$g_c'(i) = \frac{\sum_{k=iN_s}^{(i+1)N_s-1} c(k)(r'(k) - g_p'(i)r'(k-T_i))}{\sum_{k=iN_s}^{(i+1)N_s-1} c(k)c(k)} \quad (11)$$

where $c(k)$ is the decoded fixed codebook signal given by $c(k) = \sum_{i=0}^{3} s_i\delta(k-p_i)$, where $\{s_i\}$ and $\{p_i\}$ are the signs and the pulse positions decoded from the bitstream, respectively. From $r'(k)$, $g_p'(i)$ and $g_c'(i)$, we obtain the enhanced coder-specific parameters as described in Section 2.

3.3. SNR-based Feature Enhancement

At high input SNR, the speech enhancement algorithm has a tendency to degrade the speech quality. Sometimes, it causes spectral distortion for the enhanced speech. Therefore, the feature enhancement procedure described so far should be carefully applied according to the SNR of decoded speech. In other words, we only apply (7) when the estimated SNR (ESNR) is less than SNR_{thr}. Therefore, (7) is modified as

$$\hat{\omega}_{s,l} = \begin{cases} \omega_{n,l} + \mu_g(\omega_{q+e,l} - \omega_{n+q,l}), & \text{if ESNR} < \mathrm{SNR}_{thr} \\ \omega_{n,l}, & \text{otherwise.} \end{cases} \quad (12)$$

Also, we only apply the techniques described in Section 3.2 when ESNR $< \mathrm{SNR}_{thr}$. The ESNR is obtained as follows: As shown in Fig. 1, we let $s_{n+q}(l,k)$ and $s_{q+e}(l,k)$ be the k-th sample at the l-th frame of a decoded speech signal before and after applying the speech enhancement algorithm, respectively. The ESNR for a sentence is estimated as

$$\mathrm{ESNR} = \frac{10}{L}\sum_{l=1}^{L}\log_{10}\frac{\sum_{k=0}^{N-1} s_{q+e}^2(l,k)}{\sum_{k=0}^{N-1}(s_{n+q}(l,k) - s_{q+e}(l,k))^2} \quad (13)$$

where L is the total number of frames in the sentence and N is the frame size. In this measure, we assume $s_{q+e}(l,k)$ be a clean speech signal due to speech enhancement. We set SNR_{thr} = 40 dB from preliminary experimental results. By doing this, the feature enhancement is selectively applied to the noisy speech whose input SNR is below 30 dB or more.

4. SPEECH RECOGNITION EXPERIMENTS

We evaluate the performance of the proposed feature enhancement algorithm by using a connected digit recognition task, and present the recognition performance of the spectral envelope parameter enhancement algorithm, the coder-specific parameter reestimation algorithm, and the SNR-based feature enhancement algorithm. The HMM structure and subword models were the same as in [13]. Each digit was modeled by a set of left-to-right continuous density HMM's. In this task, we used a total of 274 context-dependent subword models, which were trained by maximum likelihood estimation. Subword models contained a head-body-tail

Table 1. Comparison of word accuracies (%) between the bitstream-based front-end and the conventional wireline and wireless front-ends for connected digit strings

Front-end	SNR (dB)				
	0	10	20	30	∞
Wireline	26.1	71.4	90.3	95.9	96.2
Wireless	30.1	56.8	87.8	92.6	94.8
Bitstream	22.3	71.6	92.0	95.5	96.2

Table 2. Word accuracies (%) of the bitstream-based front-end with each step of feature enhancement for connected digit strings in noisy environments

Enhancement Method	SNR (dB)				
	0	10	20	30	∞
Baseline	22.3	71.6	92.0	95.5	96.2
LSP	29.6	73.1	92.2	95.6	95.9
LSP+DA	32.5	81.3	93.6	95.7	96.0
LSP+RE	31.0	78.8	92.1	94.7	95.1
LSP+DA+SNR	32.5	81.3	93.6	95.7	96.2
LSP+RE+SNR	31.0	78.8	92.1	94.7	96.2

structure. The head and tail models were represented with three states, and the body models were represented with four states. Each state had eight Gaussian mixtures. Silence was modeled by a single state with 32 Gaussian mixtures. As a result, the recognition system had 274 subword HMM's, 831 states, and 6,672 mixtures. The training set and the test set consisted of 9,766 and 1,584 digit strings, respectively, recorded over the PSTN. The length of all digit strings for the testing was 14. The recognition experiments were done with an unknown length grammar. The word recognition accuracy was computed by counting insertion, deletion, and substitution errors.

Table 1 shows the recognition accuracies of the three front-ends according to different SNR's for the connected digit strings. The bitstream-based front-end gives a comparable performance to the wireline front-end and better performance than the wireless front-end except at 0 dB SNR. At low SNR's, the recognition performance was degraded severely for all the front-ends. Table 2 shows the word accuracies of the bitstream-based front-end versus different SNR's. In the table, *baseline* means that we do not apply the feature enhancement algorithm, and thus the results are equal to the bottom row of Table 1. We first applied the spectral parameter enhancement of (7) to the noisy speech (denoted as LSP). However, the improvement of word accuracy is not much except at 0 dB SNR, and also the word accuracy under the clean condition is reduced. Next, we added the coder-specific parameter enhancement algorithm to the spectral parameter enhancement. Compared with the direct assignment (denoted as DA) method and the reestimation method (denoted as RE) method, the DA method is better than the RE method for connected digit string recognition. Except for clean speech, the feature enhancement algorithm reduces the word error rates by around 20% and 30% at 20 dB SNR and 10 dB SNR, respectively [2]. Finally, the SNR-based feature enhancement method (denoted as LSP+DA+SNR) restores the recognition performance to the baseline accuracy under the clean condition.

[2] A significance test showed that the improvement of the feature enhancement algorithm over the baseline was statistically significant.

5. CONCLUSIONS

We have proposed a feature enhancement algorithm for wireless speech recognition in adverse environments. The feature enhancement algorithm was realized by combining a conventional speech enhancement technique in the receiver side of the wireless communications system. In other words, it utilized the feature parameters obtained from the decoded speech and an enhanced version. Moreover, the coder-specific parameters such as the residual energy, the fixed codebook gain parameter, and the adaptive codebook gain parameter could be reestimated by applying the speech enhancement algorithm to the decoded residual signal. Additionally, the algorithm was selectively applied depending on the estimated SNR. We performed connected digit HMM recognition experiments to show the effectiveness of the proposed feature enhancement algorithm. By incorporating the proposed feature enhancement algorithm into the bitstream-based front-end, we could obtain the improved recognition performance in noisy environments without hurting the performance in clean environment.

6. REFERENCES

[1] J.-C. Junqua and J.-P. Haton, *Robustness in automatic speech recognition*, Boston, MA: Kluwer Academic, 1996.

[2] A. Gallardo-Antolin, *et al.*, "Avoiding distortions due to speech coding and transmission errors in GSM ASR tasks," in *Proc. ICASSP*, Phoenix, AZ, pp. 277-280, Mar. 1999.

[3] B. T. Lilly and K. K. Paliwal, "Effect of speech coders on speech recognition performance," in *Proc. ICSLP*, Philadelphia, PA, pp. 2344-2347, Oct. 1996.

[4] C. Mokbel, *et al.*, "Towards improving ASR robustness for PSN and GSM telephone applications," *Speech Communication*, vol. 23, nos. 1-2, pp. 141-159, Oct. 1997.

[5] H. K. Kim and R. V. Cox, "Bitstream-based feature extraction for wireless speech recognition," in *Proc. of ICASSP* , Istanbul, Turkey, pp. 1207-1210, June 2000.

[6] H. K. Kim and R. V. Cox, "A bitstream-based front-end for wireless speech recognition," submitted to *IEEE Trans. Speech Audio Processing*, 2000.

[7] T. Honkanen, *et al.*, "Enhanced full rate speech codec for IS-136 digital cellular system," in *Proc. ICASSP*, Munich, Germany, pp. 731-734, Apr. 1997.

[8] C.-H. Lee, *et al.*, "Improved acoustic modeling for large vocabulary continuous speech recognition," *Computer Speech Language*, vol. 6, no. 2, pp. 103-127, Apr. 1992.

[9] R. Martin and R. V. Cox, "New speech enhancement techniques for low bit rate speech coding," in *Proc. 1999 IEEE Workshop on Speech Coding*, Porvoo, Finland, pp. 165-167, June 1999.

[10] J. S. Rustagi, *Optimization techniques in statistics*, Boston, MA: Academic Press, 1994.

[11] B. Widrow and S. D. Stearns, *Adaptive signal processing*, Englewood Cliffs, NJ: Prentice-Hall, 1985.

[12] G. S. Kang and L. J. Fransen, "Application of line spectrum pairs to low bit rate speech coders," in *Proc. ICASSP*, Tampa, FL, pp. 7.3.1-7.3.4, Mar. 1985.

[13] M. Rahim, B. H. Juang, W. Chou, and E. Buhrke, "Signal conditioning techniques for robust speech recognition," *IEEE Signal Processing Lett.*, vol. 3, no. 4, pp. 107-109, Apr. 1996.