Submitted to EUROSPEECH 2003, September 1-4, 2003 - Geneva, Switzerland

Enhancement of Noisy Speech for Noise Robust Front-End and Speech Reconstruction at Back-End of DSR System

Hyoung-Gook Kim, Markus Schwab, Nicolas Moreau, and Thomas Sikora

Department of Communication Systems
Technical University of Berlin, Germany
{kim|schwab|moreau|sikora}@nue.tu-berlin.de

Abstract

This paper presents a speech enhancement method for noise robust front-end and speech reconstruction at the back-end of Distributed Speech Recognition (DSR). The speech noise removal algorithm is based on a two stage noise filtering LSAHT by log spectral amplitude speech estimator (LSA) and harmonic tunneling (HT) prior to feature extraction. The noise reduced features are transmitted with some parameters, viz., pitch period, the number of harmonic peaks from the mobile terminal to the server along noise-robust mel-frequency cepstral coefficients. Speech reconstruction at the back end is achieved by sinusoidal speech representation. Finally, the performance of the system is measured by the segmental signal-noise ratio, MOS tests, and the recognition accuracy of an Automatic Speech Recognition (ASR) in comparison to other noise reduction methods.

1. Introduction

The European Telecommunication Standards Institute (ETSI) STQ-Aurora group has created a new work item to address the standardization of an extended front-end for Distributed Speech Recognition (DSR) of tonal languages as well as speech reconstruction for DSR systems [1]. Figure 1 presents a block diagram of future ETSI Aurora standards enabling DSR of tonal languages and speech reconstruction. The noise robust features are compressed and transmitted to the server for recognition back-end processing. For speech reconstruction at back-end server, the number of harmonic peaks and pitch period are transmitted as additional parameters along with the DSR bit stream. In this paper we apply the two stage LSAHT noise filtering method to speech reconstruction at the back-end of the DSR system under various noise conditions.

2. Noise robust front-end algorithm

The speech enhancement methods in combination with feature extraction improve both speech recognition performance and the quality of speech reconstruction under noisy conditions. Usually, speech enhancement problem is addressed from the estimation point of view in which the clean speech is estimated under the uncertainty of speech presence [2] in noisy observations. The idea of utilizing the uncertainty of speech presence in the noisy spectrum has been applied by many authors to improve the performance of speech enhancement systems. In this paper, we present a simple modified log-spectral amplitude (LSA) speech estimator [2] and harmonic tunneling (HT) [3].

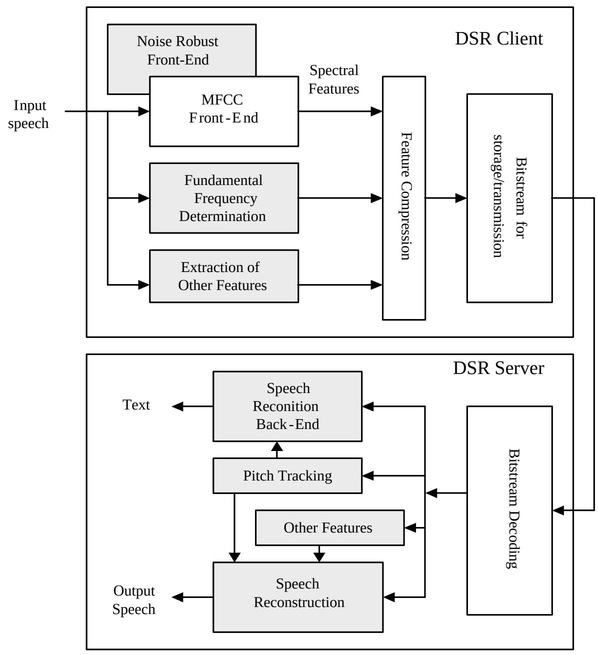

Figure 1: *Block diagram of DSR system*

2.1. Noise Reduction

A simplified block diagram of a two stage noise filtering system LSAHT based on LSA speech estimator and HT is shown in figure 2. Let $x(n)$ denote the 8 kHz sampled input speech, which is assumed to be the sum of a clean speech $s(n)$ and disturbing noise $d(n)$. The observed noisy signal $x(n)$ is divided into overlapping frames. A pre-emphasis filter is then used to emphasize the higher frequency components. In the frequency domain the short-time magnitude spectrum $A(k, i)$ of $x(k, i)$ at time frame i and frequency bin k is estimated by:

$$X(k, i) = A(k, i)e^{j\phi(k,i)} \tag{1}$$

As first noise reduction stage, the estimation of clean speech is obtained by applying a modified log-spectral amplitude gain function $G_{LSA}(k, i)$ to each spectral component of the noisy speech signal:

$$O(k, i) = G_{LSA}(k, i)A(k, i), \tag{2}$$

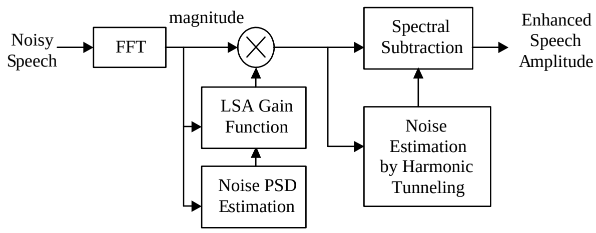

Figure 2: *Block diagram of the speech enhancement*

where $G_{LSA}(k,i)$ is derived by

$$G_{LSA}(k,i) = \frac{\xi(k,i)}{1+\xi(k,i)} \exp\Big(0.5 \int_{t=\nu(k,i)}^{\infty} \frac{e^{-t}}{t} dt\Big), \quad (3)$$

where the a posteriori signal-to-noise-ratio SNR $\gamma(k,i)$, the a priori SNR $\xi(k,i)$, and $\nu(k,i)$ are defined as:
$\gamma(k,i) = \frac{A(k,i)}{\lambda_d(k,i)}$, $\nu(k,i) = \frac{\xi(k,i)}{1+\xi(k,i)}\gamma(k,i)$, and
$\xi(k,i) = \beta G_{LSA}(k,i-1)\frac{\gamma(k,i)}{1-q(k,i)} + (1-\beta)P\{\gamma(k,i)-1\}$
using

$$P\{\gamma(k,i)-1\} = \begin{cases} \gamma(k,i)-1 & \text{if } \gamma(k,i) \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $\beta \in [0,1]$ is the SNR smoothing factor, $q(k,i)$ is an estimate of speech absence priori probability and $\lambda_d(k,i)$ is a noise spectrum estimate. The spectral gain is modified by the speech absence probability $q(k,i)$, which is estimated for each frequency bin and each frame, and which controls the update of the estimated noise spectrum when speech is present. Therefore, the noise estimate is obtained by averaging past spectral power values by using a time-varying frequency-dependent smoothing parameter. This parameter adjusted by the speech absence probability in adverse environments involving non-stationary noise, weak speech components and low input SNR:

$$\lambda_d(k,i) = \alpha_d(k,i)\lambda_d(k,i-1) + (1-\alpha_d(k,i))A(k,i) \quad (5)$$

using a smoothing parameter

$$\alpha_d(k,i) = 1 - F_d|\gamma(k,i-1)|q(k,i) \quad (6)$$

with $F_d \in [0,1]$ constant.

The a priori speech absence probability $q(k,i)$ is controlled by the result of the minimum tracking. The minimum values $M(k,i)$ of an average $E(k,i)$ of the short-time magnitude spectrum are calculated within windows of S frames whether speech is present or not. The minimum value for the current frame is found by a comparison with the stored minimum value:

$$M(k,i) = \min_{s=0..S}\{M(k,i-s), E(k,i)\}, \quad (7)$$

where the average of the short-time noisy spectrum is performed over B frames by:

$$E(k,i) = \frac{1}{B}\sum_{i=0}^{B-1} A(k,i). \quad (8)$$

The indicator function $I(k,i)$ for the voice activity detector is defined by

$$I(k,i) = \begin{cases} 1 & \text{if } A(k,i) < M(k,i)T(k,i) \\ & \quad E(k,i) < M(k,i)\tilde{T}(k,i) \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

using the thresholds $T(k,i) = 1 + 4\exp\Big[-G_{LSA}(k,i-1)\Big]$ and $\tilde{T}(k,i) = 1 + 0.5\exp\Big[-G_{LSA}(k,i-1)\Big]$. The a priori probability for speech absence is then obtained by

$$q(k,i) = \alpha_q q(k,i-1) + (1-\alpha_q)I(k,i), \quad (10)$$

where $\alpha_q \in [0,1]$ is the time-smoothing factor. Although this LSA estimator proved very efficient in reducing musical noise phenomena, low bit rate speech coders are very sensitive to remaining residual background noise after LSA estimation. Therefore, a second noise reduction stage is employed. From the magnitude spectrum $O(k,i)$ out of the first noise reduction, the voicing level is obtained by normalizing spectral autocorrelation at a lag equal to a pitch period in frequency domain. At the next stage, the peak detector is used to find the number of peaks and the frequency bin of the peak corresponding to the highest harmonic within the auto-correlation. Each of these candidate peaks is analyzed to categorize it as a peak coming from either a voiced speech harmonic or noise. To determine the harmonic amplitude $O(h,i)$ and harmonic frequency in the frame h, we do the following

$$O(h,i) = max_{m\in[a,b]}(|O(m,i)|), \quad (11)$$

where $a = floor((harmo-c)(f_0/S_r/N))$ using the sampling rate S_r and the estimated fundamental frequency f_0, and $b = ceil((harmo+c)(f_0/S_r/N))$. $c \in [0,0.5]$ determines the tolerated non-harmonicity. The estimate $\lambda_{HT}(k,i)$ of the noise is then obtained by sampling the noise spectrum in the tunnels between the harmonic spectral peaks and by interpolation of the frequency and time from the adjacent noise spectra in the surrounding tunnels. Finally, the enhanced spectral amplitude $\tilde{S}(k,i)$ is achieved by spectral subtraction:

$$\tilde{S}(k,i) = O(k,i) - \lambda_{HT}(k,i). \quad (12)$$

2.2. Mel cepstrum feature vector extraction

The noise-reduced power spectrum $\tilde{S}(k,i)$ in the frequency range between 64 Hz and 4000 Hz is then Mel-filtered. The low-frequency components are ignored. The rest of the frequency range is warped into a Mel-frequency scale using the equation

$$Mel(k,i) = 2595\log_{10}\Big(1+\frac{k}{700}\Big). \quad (13)$$

The Mel-frequency scale range is divided into 23 equal-sized, half-overlapping bands. The output of the mel-filter is the weighted sum of the noise-reduced magnitude spectrum in each band as follows:

$$F(m,i) = \sum_{k=b_m-\delta_m}^{b_m+\delta_m} \tilde{S}(k,i)U_{\delta_m}(k+b_m) \quad (14)$$

where each band wide $\delta_m = \delta_{m-1}$, the center of each band $b_m = b_{m-1} + \delta_m$, and triangular filter

$$U_{\delta_m}(k) = \begin{cases} 1-|k|/\delta_m & \text{if } |k| < \delta_m \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

are applied. The 23 log-spectral values of mel-filtering are then subjected to a natural logarithm function. With a 23-point Discrete Cosine Transformation, 13 cepstral coefficients are calculated as follows:

$$C(\kappa,i) = \sum_{m=1}^{M} \log\{|F(m,i)|\}\cos\Big[\kappa(m-0.5)\frac{\pi}{M}\Big], \quad (16)$$

where M is the number of triangular filter, κ is the number of cepstrum coefficients.

For a low bit rate speech compression and decompression we use the methods in [4] and [5]. The pitch period value is quantized using 7 bits. The number of harmonic peaks is quantized using 3 bits for the past, current and future frames. The MFCC feature vectors are quantized using a 4-split VQ with 37 bits. The streams of the compressed MFCC feature vectors, the compressed pitch period value, and the compressed number of harmonic peaks are multiplexed together to form the output bit stream for storage or transmission.

2.3. Reconstruction of speech at back-end server

The transmitted bit stream to server is fed into a stream of compressed MFCC feature vectors, a stream of compressed pitch, and a stream of number of harmonic peaks. The decompressed MFCC feature vectors may be used by the speech recognition back-end. For the speech reconstruction, the MFCC feature vectors are transformed back into the Mel-frequency domain by inverse DCT and the spectral magnitude is computed by exponentiation from the log-spectra [4]:

$$\breve{S}(k,i) = \exp\Big(-\frac{2}{M}\sum_{s=1}^{M} C(s,i)\cos\Big[\frac{(2m+1)s\pi}{2M}\Big]\Big). \quad (17)$$

Speech is synthesized using a harmonic sinusoidal model from the decompressed MFCC feature vectors, the decoded pitch values, and the number of harmonic peaks for voicing decision by

$$\tilde{s}_i(j) = \sum_{l=0}^{L-1} \bar{S}_l(j)\cos\Big(\tilde{\phi}_l(j)\Big), \quad (18)$$

where the speech sample $\tilde{s}_i(j)$ is synthetisized as the sum of a number of harmonically related sinusoids with amplitude $\bar{S}_l(j)$ at multiples of the fundamental frequency and synthetic phase $\tilde{\phi}$. For voiced speech, the model is based on the assumption that the perceptually important information resides mainly in the harmonic samples of the pitch frequency. Because of the relatively slow variation in the amplitude between successive frame and the insensitivity of the human auditory system to slight inconsistences in the speech amplitude, a straight forward linear interpolation is given by

$$\bar{S}_l^i(j) = \breve{S}_0^i \cdot j + \Big(\breve{S}_l^{i+1} - \breve{S}_l^i\Big)\Big(\frac{j}{L}\Big). \quad (19)$$

The phase is reconstructed from the decoded pitch values using a quadratic model which assumes linear pitch variations:

$$\tilde{\phi}_i^l(j) = l f_0^{i-1} j + \frac{l(f_0^i - f_0^{i-1})}{2N} j^2 + \varphi_l, \quad (20)$$

where f_0^{i-1}, f_0^i are the pitch frequency values for the $(i-1)^{th}$ frame and the i^{th} frame respectively, N is the frame size in samples, and φ_l is zero for harmonics below a threshold frequency called voicing and a random variable uniformly distributed in $\in [-\pi, \pi]$ for harmonics above the voicing frequency. For unvoiced speech, the magnitude spectrum is sampled at 100 Hz and a uniformly distributed random phase is applied to each frequency component.

3. Experimental Results

The performance of the proposed algorithm is measured using segmental SNR improvement in speech segments, recognition accuracy improvement, subjective study of speech spectrograms, and listening test.

3.1. Segmental SNR improvement

To measure the performance of the proposed algorithm in comparison to other one-channel noise reduction methods, the segmental signal-to-noise ratio ($segSNR$) at back-end of DSR is computed by $SNR_{improve} = segSNR_{out} - segSNR_{in}$ for the enhanced speech signals at back-end of DSR. Three types of background noise - white noise, car noise and factory noise - were artificially added to different portions of the data at SNR of 5 dB and -5 dB. Table 1 shows that LSAHT algorithm gives best results for input SNR 5 dB and -5 dB compared to the results of PSS, MS, DLSA and NSMR.

Table 1: *Comparison of segmental SNR improvement of different one-channel noise estimation methods. PSS: Power Spectral Subtraction, MS: spectral subtraction based on minimum statistics [6], DLSA: log-spectral amplitude speech estimator by spectral minimum tracking*
[7], *NSMR: the ratio of the spectral amplitude of the noisy speech to its minimum* [8] *and LSAHT: the proposed noise reduction method using two stage noise filtering.*

methods	Input SNR [dB]					
	white		car		factory	
	5	-5	5	-5	5	-5
PSS	4.3	7.3	5.3	8.1	4.1	7.3
MS	7.8	12.3	8.4	13.5	7.4	11.9
DLSA	7.9	12.6	8.6	13.2	7.2	12.1
NSMR	8.9	13.6	9.1	13.3	8.5	12.7
LSAHT	9.1	14.9	11.3	15.7	10.0	14.3

3.2. Recognition accuracy in a DSR system

For evaluation of the improvement of speech recognition with presented front-end, the Aurora 2 database together with a hybrid HMM/MLP ASR system (351 inputs, 420 hidden units and 24 outputs) using forward-backward training algorithm [9] have been chosen and two training modes are used: training on clean data and multi-condition training on noisy data. The feature vector consists of 39 parameters: 13 mel frequency cepstral coefficients plus delta and acceleration calculations. The mel-cepstrum coefficients are fed to the MLP (multi-layer perceptron) for the non-linear transformation consisted of 9 frames. The proposed LASHT-filtering front-end was compared to a NSMR front-end, LSA front-end, and MS front-end. For the noisy speech results, we averaged the word accuracies between 0 dB and 20 dB SNR. In the table 2, set A, B, and C refer to matched noise condition, mismatched noise condition, and mismatched noise and channel condition, respectively. Table 2 describes the results of the recognition accuracy.

As seen in the results of table 2, LSAHT provides much better performance than DLSA front-end, MS front-end, and NSMR front-end.

3.3. Speech spectrograms and listening test

In order to visualize the effect of the noise reduction algorithm based on LSAHT, the spectrograms of noisy speech and the reconstructed speech at back-end server are shown in figure 3.

Table 2: *Comparisons of word accuracies (%) between several front-ends on the Aurora 2 database*

Training Mode	Set A	Set B	Set C	Overall
Multicondition	86.91	86.61	86.66	86.73
Clean only	72.34	72.70	86.62	77.22
Average	79.63	79.65	86.64	81.97

(a) Word accurancy of DSLA front-end

Training Mode	Set A	Set B	Set C	Overall
Multicondition	89.92	88.41	86.86	88.40
Clean only	74.16	73.01	82.13	76.43
Average	82.04	80.01	84.50	82.42

(b) Word accurancy of MS front-end

Training Mode	Set A	Set B	Set C	Overall
Multicondition	89.65	88.35	86.88	88.29
Clean only	79.28	78.82	82.13	80.08
Average	84.47	83.59	84.51	84.19

(c) Word accurancy of NSMR front-end

Training Mode	Set A	Set B	Set C	Overall
Multicondition	91.45	90.21	89.13	90.26
Clean only	84.32	82.41	82.78	83.17
Average	87.89	86.31	85.96	86.69

(d) Word accurancy of LSAHT front-end

The noisy spectrograms in the upper image of figure 3 was recorded in a busy street with a SNR of about 5 dB. The spectrogram of the reconstructed speech at back-end server is depicted in the lower parts of figure 3. Dark gray areas correspond to the speech components while background noise is light gray. The picture clearly indicates that only speech portions pass the system whereas the noise is suppressed. To evaluate the quality of four (MS,DLSA, NSMR, LSAHT) speech enhancement methods of DSR back-end speech synthesizers, a subjective Mean-Opinion-Score (MOS) was performed with noisy speech corrupted by car noise at SNR 10 dB. The noisy uncoded speech scored 2.16. The MS, the DLSA and, NSMR and LSAHT back-end synthesizer scored 2.53, 2.43, 2.65 and 2.83 respectively.

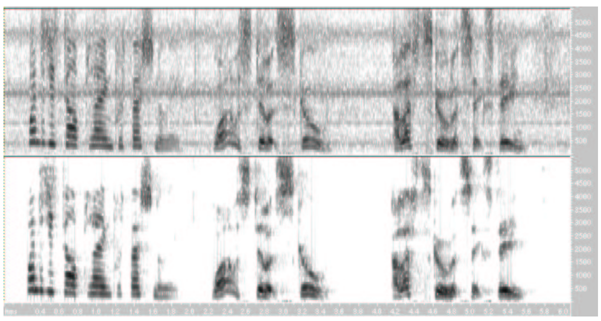

Figure 3: *Spectrograms of noisy speech, reconstructed speech at back-end of DSR system.*

4. Acknowledgements

Hyoung-Gook Kim would like to thank Cortologic AG for the experiments on the Aurora 2 database.

5. References

[1] ETSI DSR Applications and Protocols Working Group, “New Aurora activity for standardization of a front-end extension for tonal language recognition and speech reconstruction”, June 2001.

[2] Malah, D., Cox, R. V., AccardiLee, A. J., “Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments”, in Proc. ICASSP, Phonix, AZ, vol. 1, pp. 201–204, March 1999.

[3] Ealey, D., Kelleher, H., Pearce, D., “Harmonic tunneling: tracking non-stationary noises during speech”, in EUROSPEECH, Aalborg, pp. 437–410, Sep. 1999.

[4] Ramabadran, T., Meunier, J., Jasiuk, M., and Kusher B.,“Enhancing Distributed Speech Recognition with back-end speech reconstruction”, in EUROSPEECH, pp. 1859–1862, Sep. 2001.

[5] Chazan, D., Cohen, D. G., Hoory, R., Zibulski, M., “Low bit rate speech compression for playback in speech recognition systems”, in EUSPICO, pp. 1281–1284, Sep. 2000

[6] Martin, R., “Spectral subtraction based on minimum statistics”, in EUSPICO, pp. 1182–1185, Sep. 1994.

[7] Doblinger, G., “Computationally efficient apeech enhancement by spektral minimum tracking in subbands”, in EUSPICO, pp. 1513–1516, Sep. 1995.

[8] Kim, H.-G., Ruwisch, D., “Speech enhancement in non-stationary noise environment”, in ICLSP, pp. 1829–1832, Sep. 2002.

[9] Hennbert, J., Ris, C., Bourlard, H., Renals, S., Morgan, N., “Estimation of global posteriors and forward-backward training of hybrid HMM/ANN Systems”, in EUROSPEECH, pp. 1951–1954, Sep. 1997