ON THE USE OF WIDEBAND SIGNAL FOR NOISE ROBUST ASR

Dušan Macho and Yan Ming Cheng

Human Interface Lab, Motorola Labs
1301 E. Algonquin Rd., Schaumburg, IL 60196, USA
dusan.macho@motorola.com, ycheng@labs.mot.com

ABSTRACT

Wideband audio signal will be commonly used in the near future telecommunication applications. In quiet environments, the speech recognition performance increases when using wideband signal instead of narrowband signal. For practical ASR systems, however, we are interested in whether we can benefit from wideband signal when recognizing noisy speech. Wideband speech signal, with respect to the narrowband speech signal, contains high frequency components, which usually have low intensity and, thus, are vulnerable to noise distortions. The robustness of the wideband feature set may be worse than that of the narrowband feature set, if the added high-frequency components are distorted.

We investigate whether the addition of information from high frequencies into the ASR feature set can improve the recognition performance of noisy speech. The differences between low- and high-frequency parts of the wideband speech spectrum suggest a separate processing of these two parts. We propose an algorithm in which the separate processing scheme permits us to reuse the noise robust front-end originally designed for narrowband signal. A low complexity processing is designed for high frequency components, which usually bear less information. The high-frequency information is added to the narrowband speech features in a form of de-noised filter-bank energies. The energies are appended after or before computing the cepstral features. In the best case, we obtained 13.96% average relative improvement when recognizing wideband noisy speech with respect to the narrowband noisy speech performance. The proposed algorithm is part of the recently adopted ETSI standard for the advanced front-end for distributed speech recognition.

1. INTRODUCTION

In most of today's ASR systems over telecom network, the frequency range of audio signal is limited to 0-4kHz. The use of wider frequency range (wideband signal) in telecommunication applications is expected soon. Already now, the distributed speech recognition (DSR) scheme offers the possibility of using the wideband ASR features over a narrowband network. Indeed, if considering a quiet environment, the recognition performance of an ASR system will improve when the ASR feature set includes also the information from frequencies above 4kHz.

The last assertion, however, is less obvious when considering noisy speech, which is common in telecommunication. To explain why, we use a frequency range partition as it is shown in Figure 1. We consider that the wideband (w-b) signal ranges from 0 to 8kHz. Narrowband (n-b) signal ranges from 0 to 4kHz; also, we refer to these frequencies as low frequencies (l-f). On the other hand, high frequencies (h-f) extend from 4 to 8kHz.

According to speech research [1], the intensity of typical l-f speech sounds (usually voiced speech) decreases by an average rate of 6dB/octave along the frequency axis having a small portion of their intensity in the h-f part of spectrum. On the other hand, the intensity of typical h-f speech sounds (usually noise-like sounds) is low relatively to the intensity of l-f speech sounds. As a consequence, we can observe from a w-b spectrogram in Figure 2 that the intensity of speech in the h-f part of speech spectrum is much lower than the intensity of speech in the l-f part of spectrum. Additionally, the occurrence of speech along the time axis is lower in the h-f part of spectrum than in the l-f part of spectrum.

Due to the above-mentioned properties, the h-f part of speech spectrum is more prone to noise distortions than the l-f part. In the presence of noise, the combination of the l-f speech information, which is high-SNR, with the h-f speech information, which is low-SNR, would cause that the w-b speech features become more affected by noise than the n-b speech features. Of course, in this case, the recognition performance decreases. Therefore, the benefit of adding the h-f speech information into the ASR speech representation is questionable when considering noisy speech recognition.

In this paper, we investigate whether adding the h-f speech information can improve the speech recognition performance in noisy conditions. Some recent works deal with recognition of noisy wideband speech [2], [3]. Here, we examine three ways of including the high frequency information into the ASR speech features. We use the advanced front-end [4], which was recently standardized by ETSI as a feature extraction approach for DSR in noisy environments [5]. One of the studied approaches forms part of this standard.

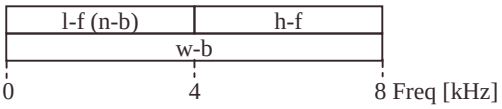

Figure 1 Partition of wideband frequency range.

2. INVESTIGATED APPROACHES

In many noise robust approaches, it is desirable to distinguish safely between the speech and non-speech portions of signal (e.g. for the noise spectrum estimation). Computationally efficient approaches for speech/non-speech separation are based on signal's energy contour. Figure 2(b) shows such a contour for the w-b signal from the spectrogram in Figure 2(a). We can observe that in the w-b energy contour of the high SNR version of the signal (dashed line; close-talking micro), both l-f and h-f speech portions can be safely identified in the background noise. In the case of low SNR version of the same signal energy contour (solid

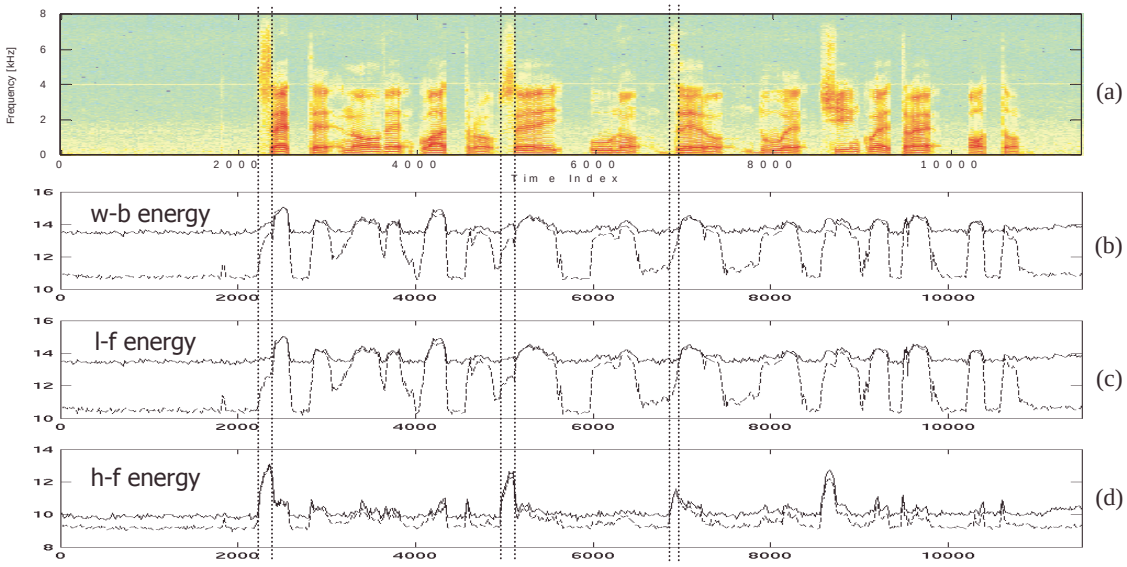

Figure 2 Wideband spectrogram of Italian digit sequence "sette nove quattro sei uno zero due cinque tre otto" - close-talking microphone (a). Corresponding wideband, low-frequency and high-frequency energy contours (b)-(d) of clean (dashed) and noisy (solid) signal.

line; hands-free micro), however, the separation of the low-energy speech parts, including both l-f and h-f speech, is very difficult. For l-f speech sounds, a periodicity indicator may help to distinguish between speech and noise. For h-f sounds, we calculated the energy contour from the h-f part of signal only (Figure 2(d)). When comparing Figures 2(b) and 2(d), we can see that the identification of the h-f speech sounds in the case of low-SNR signal (solid lines) can be done much better from the h-f energy contour than from the w-b energy contour (see e.g. the sounds selected by the vertical dashed lines). Notice that the SNR observed in the h-f energy contour is quite good because of the low-frequency character of the in-car noise in this file.

The differences and relations we observed between the l-f and h-f parts of speech spectrum up to this point, which are consequences of the nature of speech production, suggest that we could benefit from a separate processing of the l-f and h-f components of w-b signal. Actually, the idea of separate processing offers some further attractive advantages: a) the noise robust front-end designed originally for the n-b signal can be reused for processing of the l-f part of w-b signal, and b) due to a relatively lower information content, the processing of the h-f part of w-b signal may be less complex than that of the l-f part.

In this paper, we investigate three ways of using the h-f information for noisy speech recognition. In each case, speech features are Mel filter-bank based cepstral coefficients.

2.1. Description of approaches

The first approach, which is the most straightforward one, does not make distinction between the l-f and h-f parts of w-b signal. The basic modification with respect to a n-b front-end is the increase of sampling frequency, which is reflected in the respective parts of the front-end such as number of samples of the analysis window and its shift, FFT order, and possibly, number of filter-bank bands.

The other two approaches process the l-f and h-f parts of w-b signal separately. Basic concept is depicted in Figure 3. Both l-f and h-f parts of w-b signal are obtained by filtering the input w-b signal by a couple of quadrature mirror filters. Decimation by 2 is applied to both filtered signals. An efficient and complex noise reduction scheme of the n-b robust front-end (based on a 2-stage Mel Wiener filter) is reused for the l-f part of w-b signal. Usual Mel cepstrum processing steps (log filter-bank energy estimation and discrete cosine transform) are applied on the de-noised l-f signal to obtain cepstral features. The noise reduction for the h-f part of signal is computationally much simpler because it is applied on a small number of Mel-spaced filter-bank energies (e.g. 2 or 3). We use a linear spectral subtraction (SS) in this case.

In the first of the two approaches treating the l-f and h-f spectrum separately, the de-noised h-f log filter bank energies are appended to the cepstral features obtained from the l-f processing path. In this way, the size of speech feature vector increases by the number of appended log filter-bank energies.

In the second approach, the de-noised h-f log filter-bank energies are appended to the l-f log filter bank energies prior to the cepstrum calculation. In this way, cepstral features devoted to the both l-f and h-f representations of w-b signal are optimally balanced and the final number of speech features is the same as in the n-b robust front-end. Notice, however, that l-f and h-f log filter-bank energies come from rather different processing steps – as a consequence we found the h-f filter-bank energies are mismatched in terms of the intensity scale with the l-f filter-bank energies. To minimize this mismatch, we apply an adjustment to the h-f filter-bank energies, which is based on an encoding/decoding scheme. Both spectral subtraction and the encoding/decoding scheme are described in the following sections.

2.1.1. Spectral subtraction in h-f

Linear spectral subtraction is applied on the h-f filter-bank energies $E_h(k)$ like

$$E_{SS\_h}(k) = \max\left\{E_h(k) - \alpha \cdot \hat{N}_h(k),\ \beta \cdot E_h(k)\right\}, \quad 1 \le k \le K_h \qquad (1)$$

where K_h is the number of bands, α=1.5 and β=0.1 are the overestimation and flooring factors, respectively. All K_h, α and β were set empirically. The noise estimation $\hat{N}_h(k)$ is updated by

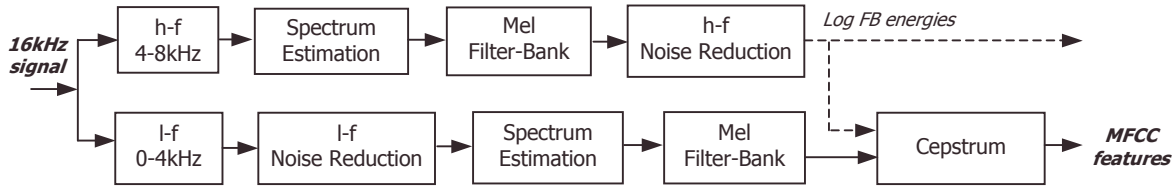

Figure 3 Separate processing of low-frequency and high-frequency components of wideband signal.

using only the frames labeled by voice activity detector as noise. An energy-based voice activity detector is used, where the current frame is labeled as speech if the difference between the current frame log energy and the long-term estimate of non-speech log energy exceeds a defined threshold.

2.1.2. Encoding/decoding scheme in h-f

The objective of the encoding/decoding (E/D) scheme is to preserve the integrity across the Mel filter-bank energies resulting from both the l-f and h-f signal processing. The E/D scheme is applied on the log version of h-f filter-bank energies $S_h(k) = \ln(E_h(k)),\ \ 1 \le k \le K_h$, and it is dependent on the l-f noise reduction. In the first step (encoding), a code is generated for each of the h-f log filter-bank energies $S_h(j)$ like

$$Code(k, j) = S_{l\_aux}(k) - S_h(j),\ \ 1 \le k, j \le K_h, \qquad (2)$$

where $S_{l\_aux}(k)$ are K_l=K_h=3 auxiliary log filter-bank energies from the 2-4kHz frequency range of the l-f signal before applying the l-f noise reduction. These auxiliary energies are available as a by-product from the original l-f noise reduction block. Such a code "memorizes" the integrity relationship between the filter-bank energies from both the l-f and h-f signals before performing the l-f noise reduction.

In the second step (decoding), a "de-noised" set of h-f log filter-bank energies $S_{code\_h}(k)$ is obtained by using the code generated previously by (2) and the following equation:

$$S_{code\_h}(k) = \sum_{j=1}^{K_h} w_{code}(j)\left(S_{den\_l\_aux}(j) - Code(j,k)\right),\ \ 1 \le k \le K_h \quad (3)$$

where $S_{den\_l\_aux}(k)$ are filter-bank energies of the de-noised l-f spectrum, and they are aligned in frequency with $S_{l\_aux}(k)$. $w_{code}(j)$ are frequency-dependent weights and their sum equals 1.0.

The h-f log filter-bank energies resulting from the E/D scheme (equation (3)), can be viewed as noise reduced versions of the original energies $S_h(k)$, assuming that all of the filter-bank energies involved in the calculation have the same SNR. Of course, in real world this assumption reflects the truth only up to a certain degree.

2.1.3. Integration of SS and E/D scheme in h-f

As discussed previously, the E/D scheme is used to minimize the intensity mismatch between the l-f and h-f filter-bank energies. The mismatch was introduced by the different noise reductions used in the l-f and h-f parts of w-b signal. In practice, the h-f energies from spectral subtraction $S_{SS\_h}(k)$ are adjusted as follows:

$$S_h(k) = \lambda \cdot S_{code\_h}(k) + (1-\lambda) \cdot S_{SS\_h}(k),\quad 1 \le k \le K_h \qquad (4)$$

where λ = 0.7 was determined experimentally.

Cepstral features corresponding to the w-b signal are calculated from log filter-bank energy vectors that are formed by appending the h-f log filter-bank energies from (4) to the de-noised l-f log filter-bank energies. Also, the energy parameter accompanying cepstral features is computed by using the de-noised version of both the l-f and h-f parts of the w-b spectrum.

3. RECOGNITION EXPERIMENTS

3.1. Testing scenario

We used four SpeechDat Car (SDC) digit databases: Spanish, Finnish, Danish and Italian, a subset of the database set [6] distributed by ETSI for the Aurora standardization. Databases contain both 8 and 16kHz data and they were collected in car under different driving conditions with both close talking and hands-free microphones. Three recognition experiments were carried out for each language with different levels of mismatch between training and testing conditions: well-matched (WM), medium mismatched (MM) and highly mismatched (HM).

SDC databases contain long non-speech portions of signal at the beginning/end of each recording. Long non-speech portions tend to increase the number of insertions during recognition of noisy speech. To avoid these insertions, a voice activity detector (VAD) is usually used and the frames classified as non-speech are dropped out from recognition (see [4]). In tests presented in this paper, we use an "ideal" VAD for frame dropping to minimize the influence of VAD on the performance. The HTK Viterbi alignment was used to get the utterance boundaries in clean version of each file. Due to the correspondence between clean and noisy files, the clean speech utterance boundaries can be used also for noisy files. The ideal VAD preserves 200ms of signal before the beginning and after the end of each utterance and drops the remaining signal.

For feature extraction, we used both the standard MFCC front-end [7] and the advanced front-end [4]. Delta and delta-delta features are appended to the static features. As for the back-end configuration, the digit models have 16 states with 3 Gaussians per state. The silence model has 3 states with 6 Gaussians per state. Also, one-state short pause model is used and is tied with the middle state of the silence model.

3.2. Experiments and results

We compare recognition performances obtained on 8kHz and 16kHz data. In the first set of experiments, no separation to l-f and h-f processing was done. Simply, the sampling frequency of the front-end was modified from 8 to 16kHz: analysis window size changed from 200 to 400, window shift changed from 80 to 160, and FFT order changed from 256 to 512. We tested two different numbers of filter-bank bands, 23 and 30, but the number of static cepstral coefficients was kept the same, 12 + log energy. The ETSI standard MFCC front-end was used in these experiments. Table 1 shows the word accuracies we obtained for both the 8 and 16kHz data. We can observe an average relative improvement of 4.59% for the 16kHz signal with respect to the 8kHz signal (see the last column of Table 1) when using 23 filter-

bank bands. Particularly, improvements can be observed in WM and MM conditions, however, there is a degradation in the noisy HM test. Similar results can be observed for the 16kHz signal when using 30 filter-bank bands.

In the following experiments, the l-f and h-f parts of w-b signal were processed separately. Recognition performances are shown in Table 2 – we used the advanced front-end, thus the 8kHz baseline is much higher than in the previous tests. In the first approach, we tested the addition of the de-noised h-f log filter-bank energies to cepstral features. The best results were obtained by adding two h-f log energies. Therefore, the number of static parameters increased from 13 to 15. A slight average relative improvement of 2.25% was obtained by using this approach for 16kHz data in comparison to the performance with 8kHz data. Small improvements can be observed for WM and MM tests, while the average performance in HM test slightly decreased. The results from this test coincide with the results reported in [2] where author used a slightly modified implementation of this approach.

In the second approach, three de-noised h-f log filter-bank energies were added to the 23 de-noised l-f log filter-bank energies and this way formed log filter-bank energy vector was used to compute 12 cepstral coefficients and the log energy parameter (i.e. 13 static parameters). The results from this test are reported at the third line of Table 2. A significant average improvement of 13.96% can be observed by using this approach for w-b signal with respect to the n-b baseline (the fist line of Table 2). The performance is significantly better than what we obtained by appending the h-f log filter-bank energies to cepstral features in the previous test. Actually, in this case, the added h-f log filter-bank energies are de-correlated by the following DCT transform (but not in the previous test). Use of correlated features for diagonal covariance HMMs usually decreases the performance. The latest approach outperforms the other two techniques we evaluated and it is also used in the ETSI DSR advanced front-end standard.

4. CONCLUSIONS

Adding the information from signal frequencies above 4kHz to a narrowband ASR feature set improves the recognition performance of clean speech. In this paper, we investigated whether we can obtain similar improvement for noisy speech. After observing significant differences between the low-frequency (0-4kHz) and high-frequency (4-8kHz) portions of speech spectrum, we proposed separate processing of these two components of wideband signal to gain better noise robustness. This approach led also to further benefits: we could reuse the noise robust front-end originally designed for narrowband signal, and we could design a low complexity algorithm for high frequencies.

The high-frequency information was added to the narrowband speech features in a form of de-noised filter-bank energies. We have experimentally found that appending these energies to the narrowband filter bank energies provides a performance superior to other investigated approaches. We obtained 13.96% average relative improvement when recognizing wideband noisy speech with respect to the narrowband noisy speech performance. This indicates that a positive contribution can be achieved from the high frequency components of wideband signal when recognizing noisy speech.

5. REFERENCES

[1] Rabiner, L.R., Schafer R.W., Digital Processing of Speech Signals. Prentice-Hall, New Jersey, 1978.

[2] Hirsch, H.-G., *"Ericsson Proposal for a Robust Feature Extraction as Input to STQ-DSR WI008"*, ETSI STQ Aurora Document AU/287/01, Jan 2001.

[3] Nadeu, C., Tolos, M., *"Recognition Experiments with the SpeechDat Car Aurora Spanish Database using 8kHz- and 16kHz-Sampled Signals"*, ASRU 2001 Italy.

[4] Aurora doc. *"Motorola - France Télécom - Alcatel Advanced Front End Proposal"*, Adopted by ETSI for DSR advanced front-end evaluation, Jan 2002. Also appeared in ICSLP'02.

[5] ETSI Telecom Standards News, *"ETSI Selects Advanced Algorithm for Distributed Speech Recognition"*, available from http://www.etsi.org/pressroom/Previous/2002/STQ-Aurora.htm, Feb 2002.

[6] Aurora documents AU/225/00, AU/237/00, AU/271/00 AU/378/01 describing data and baseline results for Finnish, Italian, Spanish and Danish databases, respectively.

[7] ETSI standard doc. *"Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*, ETSI ES 201 108 V1.1.2 (2000-04), available from http://pda.etsi.org/pda/queryform.asp, April 2000.

Sampling Frequency	SP			FI			IT			DA			Average of Abs			Aver of Rel
	WM	MM	HM	WM	MM	HM	WM	MM	HM	WM	MM	HM	WM	MM	HM	
8k, baseline	92.94	83.31	51.55	92.74	80.51	40.53	92.26	73.39	51.76	87.28	67.32	39.37	**91.31**	**76.13**	**45.80**	--
16k, fb 23	93.76	84.00	50.23	93.69	84.54	35.02	93.59	76.71	41.29	89.19	70.18	30.81	**92.56**	**78.86**	**39.34**	**4.59**
16k, fb 30	93.95	83.29	53.68	92.78	82.63	34.52	93.13	76.67	40.24	88.63	71.35	32.74	**92.12**	**78.49**	**40.30**	**2.63**

Table 1 Word accuracy percentages for 8 kHz and 16 kHz signal using the standard MFCC FE and ideal VAD.

Sampling Frequency	SP			FI			IT			DA			Average of Abs			Aver of Rel
	WM	MM	HM	WM	MM	HM	WM	MM	HM	WM	MM	HM	WM	MM	HM	
8k, baseline	96.03	92.65	88.33	95.53	85.98	87.07	96.45	89.85	87.72	92.74	82.03	80.71	**95.19**	**87.63**	**85.96**	--
16k, 1st approach	95.68	91.63	82.70	97.19	88.10	89.54	96.67	91.73	90.89	92.35	83.33	75.73	**95.47**	**88.70**	**84.72**	**2.25**
16k, 2nd approach	95.90	93.31	87.00	97.17	90.97	88.66	97.32	92.85	89.24	93.87	85.68	78.07	**96.07**	**90.70**	**85.74**	**13.96**

Table 2 Word accuracy percentages for 8 and 16 kHz signal when using noise-robust FE and ideal VAD.