ROBUST SPEECH RECOGNITION BY PROPERLY UTILIZING RELIABLE FRAMES AND SEGMENTS IN CORRUPTED SIGNALS

Yi Chen, Chia-yu Wan, Lin-shan Lee

Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, Republic of China

chenyi@speech.ee.ntu.edu.tw, chiayui@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

In this paper, we propose a new approach to detecting and utilizing reliable frames and segments in corrupted signals for robust speech recognition. Novel approaches to estimating an energy-based measure and a harmonicity measure for each frame are developed. SNR-dependent GMM Classifiers are then trained, together with a Reliable Frame Selection and Clustering module and a Reliable Segment Identification module, to detect the most reliable frames in an utterance. These reliable frames and segments thus obtained can be properly used in both front-end feature enhancement and back-end Viterbi decoding. In the extensive experiments reported here, very significant improvements in recognition accuracies were obtained with the proposed approaches for all types of noise and all SNR values defined in the Aurora 2 database.

Index Terms—Harmonic analysis, robustness, speech recognition, Viterbi decoding.

1. INTRODUCTION

Robust speech recognition under noisy conditions has been an important yet unsolved problem. In this paper we propose a new approach, which considers the fact that even in seriously corrupted speech utterances, very often there still exist some signal frames which are reliable enough. If these reliable frames can be precisely identified or even clustered into reliable segments, they can be very helpful for recognition. Stronger voiced frames are the first candidates for such purposes, because they actually carry stronger harmonicity and higher energy. But some weak voiced and unvoiced speech frames which are reliable enough are also necessary for this purpose. This is the basic idea of this paper.

Previous works have indicated that carefully examining the characteristics of speech signals and identifying the reliability of speech information in different portions of an utterance can be helpful to many speech processing systems [1, 2]. A good example in this direction is the concept of usable speech [2-5], in which various features including pitch information were developed and integrated for extracting the usable speech segments. On the other hand, substantial efforts have been made and many approaches verified very effective for improving the speech recognition performance in noisy environments. In the category of front-end feature enhancement, good examples include feature normalization techniques such as Cepstral Mean Subtraction (CMS) [6] and Cepstral Mean and Variance Normalization (CMVN) [7], and feature transformation techniques such as PCA-based [8] or multi-eigenvector temporal filtering [9]. In these approaches accurate estimation of the statistical parameters of speech and noise signals in the utterances is the key, and correctly identifying the reliable frames and segments in the signals is certainly important. In the category of back-end processing techniques, good examples include missing data speech recognition [10-13] and weighted Viterbi decoding [14-19]. Missing data approaches consider some parts of the signals as unreliable or missing, which are thus ignored in the subsequent processing, or filled up by their optimal estimates. However, accurately identifying the missing parts in the signals remains a difficult task [10-13]. On the other hand, the concept of weighted Viterbi decoding (WVD) is that during Viterbi decoding different weights can be assigned to the acoustic scores obtained from different frames or even different feature parameters in an utterance [14-19].

The work of this paper follows the general direction mentioned above. We propose a series of approaches to detecting and utilizing reliable frames and segments in corrupted signals, including special energy and harmonicity measures, various ways to identify reliable frames/segments and approaches to using them in front-end feature enhancement and back-end Viterbi decoding. Certainly there can be infinite number of ways to realize the basic idea, and the work presented below is just one of them.

This paper is organized as follows. In section 2.1, we present an overview of the proposed approach, followed by sections 2.2-2.7 containing the details for each module. Section 3 introduces the experimental conditions, and extensive experimental results are presented in section 4. Section 5 finally makes the concluding remarks.

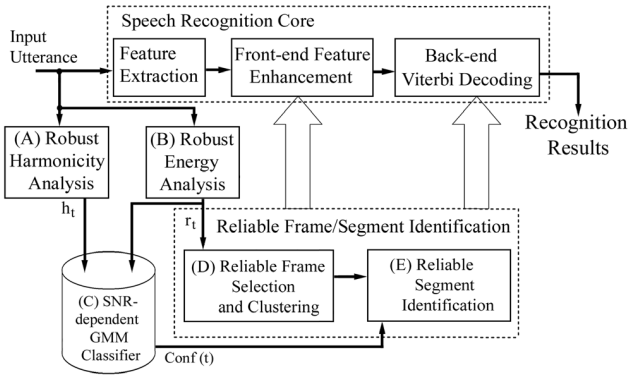

Figure 1. Overall block diagram of the proposed approach.

This work is supported by the National Taiwan University Advanced Speech Technology Scholarship.

2. PROPOSED APPROACH

2.1. Overall picture

The overall picture of the proposed approach is shown in Figure 1. The upper-most left-to-right path is the conventional robust speech recognition core: Feature Extraction followed by Front-end Feature Enhancement (e.g. feature normalization and/or transformation) and Back-end Viterbi Decoding. The approach proposed here in this paper is the lower part of the figure. We first perform Robust Harmonicity and Energy Analysis (Blocks (A)(B)) for each frame, and use the results to train the SNR-dependent GMM Classifier at the lower left corner of the figure (Block (C)). The Reliable Frame/Segment Identification then includes two parts. The Reliable Frame Selection and Clustering (Block (D)) first uses the frame energy measure to select reliable frames and cluster them into reliable segments. The Reliable Segment Identification (Block (E)) then uses the outputs of the SNR-dependent GMM Classifier to detect the most reliable frames and segments to be used.

All these results can then be properly utilized in Front-end Feature Enhancement and Back-end Viterbi Decoding. For example, if CMVN [7] is used in Front-end Feature Enhancement, the mean and variance can be estimated from those frames identified as being reliable. In Back-end Viterbi Decoding, the likelihood scores of each frame can also be weighted differently based on its reliability.

2.2. Robust Energy Analysis and Reliable Frame Selection and Clustering

We first discuss the functions of Robust Energy Analysis in Block (B) of Figure 1. For each input utterance, we first calculate the smoothed instantaneous sample energy e[n] for each signal sample, which is the energy averaged within a small window centered on the sample being considered. Here n is the sample index. We then sort all the samples in the utterance into a queue with increasing e[n], and assign a binary parameter b[n] to each sample, where b[n] is 1 if e[n] is large enough,

$$b[n] = \begin{cases} 1, & \text{if } e[n] > \mu - K * \sigma, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where μ and σ are the mean and standard deviation of all e[n] within the utterance, and K is an empirical parameter. The threshold in Eq. (1) is set automatically for each utterance, and is different from the fixed threshold used in [20]. An energy-based measure r_t is then defined for each signal frame with frame index t within the utterance, which is the average of b[n] values (0 or 1) for all the samples in the frame. Thus r_t is a real number between 0 and 1, indicating how possible a frame is reliable considering its energy behavior in the frame.

We then discuss the functions of Reliable Frame Selection and Clustering in Block (D) of Figure 1. The histogram of r_t for each utterance is first constructed with a typical example as shown in Figure 2. Note that for most frames r_t tends to either 1 or 0. A threshold T is then automatically set as the first local minimum above 0 in the histogram, as shown in Figure 2, and all frames with r_t above T are taken as first stage reliable frames. Consecutive reliable frames are then clustered. A reliable segment is obtained if the number of reliable frames in a cluster exceeds a threshold M. Isolated reliable frames or smaller clusters are simply deleted. The value of M is chosen in such a way that M frames form a segment with the minimum length of a phoneme perceivable by human auditory systems [1].

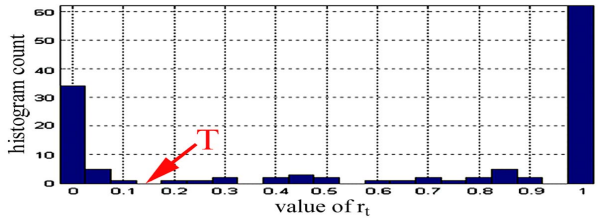

Figure 2. Histogram of r_t for a typical example utterance.

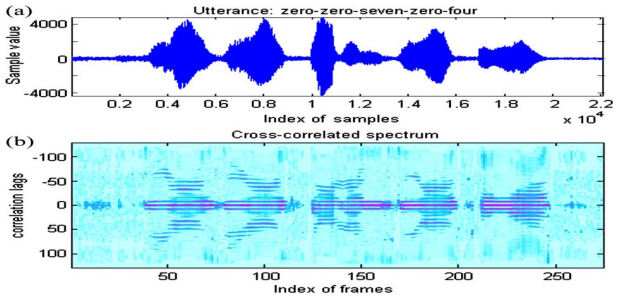

Figure 3. (a) An example utterance and (b) its cross-correlated spectra.

2.3. Robust Harmonicity Analysis

The purpose of Robust Harmonicity Analysis in Block (A) is to detect harmonic structure in the signals, since harmonic structure is a very strong indicator for voiced speech sounds.

2.3.1. Cross-correlation of frame spectra

The input frames are Hamming-windowed, low-pass filtered and transformed to the frequency domain using FFT. The magnitude spectrum of a frame is squared and cross-correlated with that of the previous frame. Harmonic structure of a frame can be enhanced with cross-correlation because of the short-term stationary property of voiced speech signals. The spectrum of the previous frame can also be considered as a "matched filter" for the current frame spectrum. Figure 3 shows an example utterance and the cross-correlated spectra of all its frames.

2.3.2. Comb-filterbank

A set of comb filters, or a "comb filterbank," is applied on the cross-correlated spectra of an utterance for robust detection of harmonic structure. A narrow Gaussian-shaped kernel function K[k], which is common to all comb filters in the filterbank, is used here to model the spreading of harmonic components in the voiced speech spectra [21, 22], as shown in Figure 4(a),

$$K[k] = \begin{cases} \exp(-k^2/\Sigma^2), & k \in [-2, 2], \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where k is the bin index inside a cross-correlated frame spectrum, and Σ^2 is a chosen constant [21, 22]. To construct a particular comb filter Comb[k, p] for a target pitch p, an intermediate filter $\overline{\text{Comb}}[k, p]$ is first defined,

$$\overline{\text{Comb}}[k, p] = \begin{cases} \sum_{\substack{m = 0, \pm 1, \pm 2, \ldots \\ mp \le N}} K[k+mp], & k = \pm 1, \pm 2, \ldots, \pm N, \\ K[1], & k = 0, \end{cases} \quad (3)$$

where p is the discrete pitch frequency, and N is the FFT order. The final comb filter Comb[k, p] is obtained by subtracting the mean of the coefficients of $\overline{\text{Comb}}[k, p]$ in Eq. (3) from all coefficients (this zero-mean property makes its response negligible

to noise with white or flat spectrum), and then normalizing all the coefficients with the vector norm of the filter coefficients such that the response is unified for all different values of p. Coefficient of Comb[k, p] for k = 0 is intentionally suppressed to suppress the weight for the zero-lag term when defining the frame harmonicity measure. The final comb filter Comb[k, p] for p = 12 is plotted in Figure 4(b). The comb filterbank then includes many such comb filters for all possible values of p for human voice.

2.3.3. Frame harmonicity

The logarithm of the cross-correlated spectra is filtered by the comb filterbank, and the outputs are half-wave rectified. Figure 5(a) is the output from the filertbank for an example utterance, where the vertical scale is the different values of p, and the horizontal scale is the frame index t. Figure 5(b) shows the same output after being sorted vertically in descending order, $\overline{Y}_t(l)$ (l = 1, 2, ..., L), where l is the order after sorting, t is the frame index, and L is the number of comb filters in the filterbank. The proposed frame harmonicity for a frame at time t, h_t, is evaluated first by the weighted sum of the sorted filterbank outputs $\overline{Y}_t(l)$ in Figure 5(b),

$$\overline{h}_t = \sum_{l=1}^{L} \alpha^{l-1} \cdot \overline{Y}_t(l) \qquad (4)$$

where α is a weighting parameter smaller than 1. Typically we set α to above 0.8 for emphasizing the largest four terms in Eq. (4). In this way, all possible pitch patterns are considered, and those having high cross-correlation with neighboring frames are emphasized. The final frame harmonicity h_t is then $\overline{h}_t$ in Eq. (4) but normalized to the range of 0 to 1 for each input utterance. The contour of h_t obtained in this way is shown in Figure 5(c).

2.4. SNR-dependent GMM Classifier

This corresponds to Block (C) in Figure 1. Given a clean speech training corpus and its transcriptions, hidden Markov models (HMMs) can be trained and used to perform forced alignment on the clean training utterances. Then voiced, unvoiced speech and non-speech frames can be located on the training utterances. If we add noise signals to these training utterances with different SNR values, for each SNR value the energy-based measure r_t defined in section 2.2 and the frame harmonicity measure h_t defined in section 2.3.3 can be calculated for each frame of these utterances. A pair of GMMs can then be trained based on the two parameters r_t and h_t, one for voiced speech frames (with strong harmonicity and relatively higher energy), and the other for unvoiced speech and non-speech frames (with weak or no harmonicity and relatively lower energy) [1], where the training frames for each class are obtained via forced alignment. In this way we have a set of very reliable SNR-dependent GMM Classifier trained for each SNR value, to be used to detect speech frames with strong voicing nature, or the nuclei of voiced phones, which are usually the most reliable parts in the noise-corrupted speech signal.

For this work based on the Aurora 2 testing environment [23], the multi-condition training set of Aurora 2 consists of utterances in five SNR conditions, i.e., clean, 20, 15, 10 and 5 dB SNR, and each utterance has a clean version in the clean training set. Therefore, for each SNR condition it is possible to train a GMM Classifier based on the stereo data from the two training sets. During testing, an SNR detector based on voice activity detection (VAD) can be used to estimate the closest SNR condition, and the classifier for 5 dB SNR is also used for all the cases with SNR lower than 5 dB. A frame is classified as a voiced speech frame

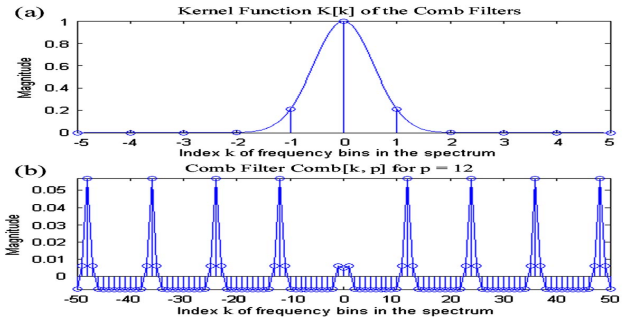

Figure 4. (a) The kernel function K[k] and (b) the final comb filter Comb[k, p] for p = 12.

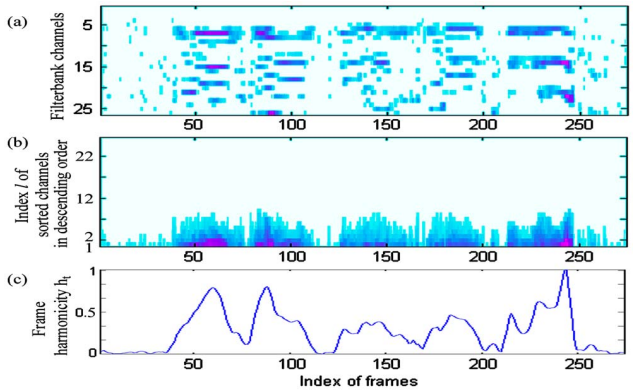

Figure 5. (a) The original and (b) the sorted output from the comb filterbank for each frame in the utterance in Figure 3, and (c) the final frame harmonicity h_t.

only if the confidence measure obtained from the ratio of the likelihood score from the GMM of voiced speech frames to that from the GMM of unvoiced speech and non-speech frames is above a threshold.

2.5. Reliable Segment Identification

In Block (E), Reliable Segment Identification, we first check to see if a reliable segment obtained from Block (D) is really reliable. This is performed based on the outputs of the SNR-dependent GMM Classifier in Block (C). A segment obtained from Block (D) is verified to be really reliable, or include reliable strong voiced frames, as long as it includes at least one frame classified as a strong voiced speech frame by the GMM Classifier. If not, the segment is deleted. This is the first step of Block (E).

The frames that are within the reliable segments verified above and also classified as speech frames by the GMM Classifier are of course confirmed as reliable frames.

The above SNR-dependent GMM Classifier is based on frame harmonicity and energy-based measures. So it can be used to reliably detect speech frames with a strong voicing nature, usually the nuclei of voiced phones or the most reliable parts in corrupted speech. However, unvoiced speech frames usually have low harmonicity values, and weak voiced speech frames usually have low energy values. They are also reliable enough due to the relatively slow-changing nature of the corrupting noise, but cannot be identified by the GMM classifier simply because they are unvoiced or weak. Fortunately, it is found that within the same reliable segment obtained via Frame Clustering, very often these two kinds of speech frames appear in the vicinity of strong voiced speech frames identified by the GMM Classifier. Therefore as the

second step of Block (E), a distance threshold D is chosen, and D frames of signals on both sides of the speech frames classified by the GMM Classifier, as long as they are within the same reliable segment identified in the above first step of Block (E), are also confirmed as reliable frames, so as to include unvoiced and weak voiced speech frames. All other frames not confirmed in this way, even if they are within the reliable segments identified by the first step of Block (E), are finally deleted. The value of D can be estimated from the statistics obtained from the noisy training set, for example the multi-condition training set of Aurora 2.

2.6. Reliable frames/segments used in front-end feature enhancement

The reliable frames/segments obtained in Blocks (D) and (E) can be properly utilized in various ways for front-end feature enhancement. A recently proposed feature enhancement front-end [20], as shown in Figure 6, is taken as a typical example of existing robust speech recognition approaches to be integrated with the approaches proposed in this paper. This front-end consists of two parts: Cepstral Mean and Variance Normalization (CMVN) for feature normalization and Two-stage PCA for feature transformation. In Two-stage PCA, a first stage PCA first transforms 14 MFCC features (C0~C12 and log-E) into 13 principal components, and in the second stage multi-eigenvector temporal filtering [9] is then performed on the temporal trajectories of these 13 principal components obtained above.

For the example four-stage feature enhancement front-end, the only changes made here are that only those frames considered as reliable in Block (D), or those identified as reliable in Block (E), are used for evaluating all the required parameters, for example the mean and variance needed for CMVN and covariance matrices needed for PCA analysis. This represents of course only one of many possible ways to use these reliable frames and segments in the front-end feature enhancement.

2.7. Reliable frame/segment information used in back-end Viterbi decoding

During Viterbi decoding, the log-likelihood scores of a feature vector for a frame at time t can be weighted by a factor w_t. The weighting factor w_t can be defined in various ways. A simple example is to have w_t dependent on the confidence measure obtained from the ratio of likelihood scores from the SNR-dependent GMM Classifier. We can further divide the evaluation of the likelihood scores in the Gaussian mixtures in HMMs into three sections, i.e. those for the original MFCC parameters, and for their first and second derivatives. The above weighting factor w_t can be used for the first section. Those used for the first and second derivatives can then be defined as below,

$$w_t^{(1)} = \sum_{k=-I_1}^{I_1} |k| \cdot w_{t+k} \Big/ \sum_{k=-I_1}^{I_1} |k|, \quad (5)$$

and

$$w_t^{(2)} = \sum_{k=-I_2}^{I_2} |k| \cdot w_{t+k}^{(1)} \Big/ \sum_{k=-I_2}^{I_2} |k| \cdot \quad (6)$$

For simplicity we can set $I_1 = 3$ and $I_2 = 2$, exactly the window sizes used respectively for defining the first and second derivatives in the Aurora 2 baseline settings [23]. Again, the above represents only one of many possible ways to use information about these reliable frames and segments in the back-end decoder.

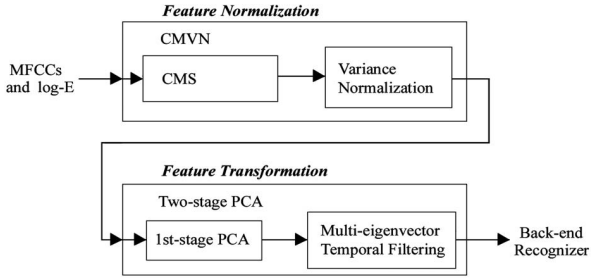

Figure 6. The feature enhancement front-end consisting of CMVN for feature normalization and Two-stage PCA for feature transformation, as a typical example of existing robust speech recognition techniques [20].

3. EXPERIMENTAL CONDITIONS

3.1. Aurora 2 database and front-end feature extraction

The experiments reported in this paper were conducted on the AURORA 2 testing environment [23], which is based on a clean speech corpus of English connected digit strings sampled at 8 kHz. Each of the two training sets, i.e. clean training and multi-condition training sets, consists of 8440 utterances. Only the clean training set was used to train the acoustic models here for tests in highly mismatched conditions. The multi-condition training set was used to train the SNR-dependent GMM Classifiers. Ten combinations of noise and channel distortions, as representatives of real-world environments and each with different SNR values, were defined in three testing sets A, B, and C and tested here. The WI007 front-end [23] gave 14 MFCC parameters (C0~C12 and log-E) as the original features for further processing, including obtaining the first and second derivatives. The HMM settings and HTK-based training and testing procedures follow the Aurora 2 specifications [23]. For tasks different from Aurora 2, a development set can be defined to play the role of multi-condition training set here.

4. EXPERIMENTAL RESULTS

4.1. Recognition performance with back-end Viterbi decoding only

We first applied the reliable frame/segment information in the back-end Viterbi decoding only. The results obtained by the proposed weighted Viterbi decoding (WVD) as mentioned in section 2.7 are in the second bar in Figure 7, as compared to the MFCC baseline of Aurora 2 in the first bar. These results are separated for different types of noise but averaged over all SNR values in Figure 7(a), for different SNR values but averaged over all types of noise in Figure 7(b), and for the three testing sets A, B, C and their average in Figure 7(c). Significant improvements were obtained in all cases. As typical examples, in Figure 7(a), the error rate reductions were 19.95% for babble noise (accuracy from 49.89% to 59.88%) in set A, and 19.79% for restaurant noise (52.59% to 61.98%) in set B.

4.2. Combination with front-end feature normalization

The results of using the reliable frames and segments obtained in Blocks (D) and (E) in front-end CMVN for feature normalization and further combined with back-end weighted Viterbi decoding are shown in Figure 8. In each set of results, the first bar is for the conventional CMVN with the mean and variance evaluated in the conventional way. The next two bars are then respectively for the results obtained with the mean and variance evaluated only from

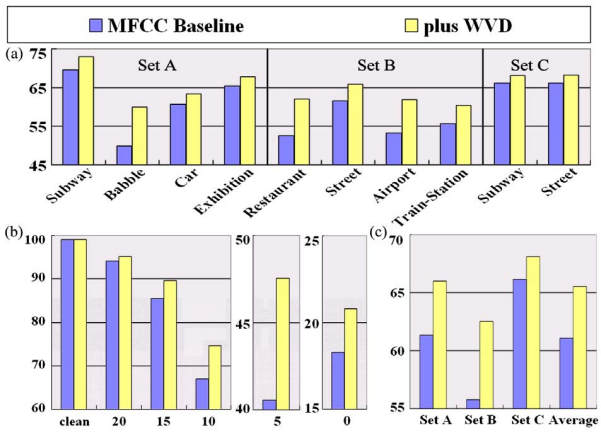

Figure 7. Comparison of recognition accuracies (%) obtained with the MFCC baseline and with weighted Viterbi decoding further applied, (a) averaged over all SNR values but separated for different types of noise, (b) averaged over all types of noise but separated for different SNR values, and (c) averaged over all SNR values and noise types but separated for sets A, B, C.

the frames selected by Block (D) ((D) Frames) or from the segments identified in Block (E) ((E) Segments). The last bar shows the results with weighted Viterbi decoding further applied (plus WVD). Clearly, very significant incremental improvements were obtained in all cases.

In Figure 8(a), the most significant improvements were obtained for car noise in set A (68.80% for CMVN to 84.73% plus WVD), airport noise (71.03% to 84.77%) and train-station noise (68.52% to 83.62%) in set B. As a good example of non-stationary noise, for airport noise the relative error rate reductions were about 32.53% (71.03% to 80.45%), 41.58% (to 83.08%), and 47.43% when Blocks (D), (E) and weighted Viterbi decoding were applied one by one in addition. In Figure 8(b), taking 10 dB as an example, the achievable accuracy was 86.29%, 88.75% and 89.72% when Blocks (D), (E), and weighted Viterbi decoding were applied one by one, corresponding to a relative error reduction of 35.74%, 47.24%, and 51.80% respectively as compared to conventional CMVN. Similar incremental improvements can be obtained in Figure 8(c) for the three sets A, B, and C, and the improvements are consistent and uniform for all three sets, with overall average improvement from 69.13% to 81.39%, which implied a relative error reduction of 39.70%.

4.3. Integration with front-end feature transformation

Here we further consider the situation that the proposed approach was applied with some existing robust speech recognition techniques, say the Two-stage PCA for feature transformation as discussed in section 2.6. The results are in Figure 9, where the first bar in each set is for conventional CMVN plus Two-stage PCA (CMVN + Two-stage PCA), i.e., using all frames to estimate the required parameters, and the rest with the proposed approaches applied one by one in addition.

In Figure 9(a) with the proposed approaches applied one by one, significant improvements were obtained for all types of noise over the original front-end. When Two-stage PCA was applied based on the reliable frames and segments identified in Block (D) or Block (E) in Figure 1, accuracies for all types of noise were successfully improved stage by stage to over 82% or more, among which the case of babble noise in set A (82.73%) is the lowest.

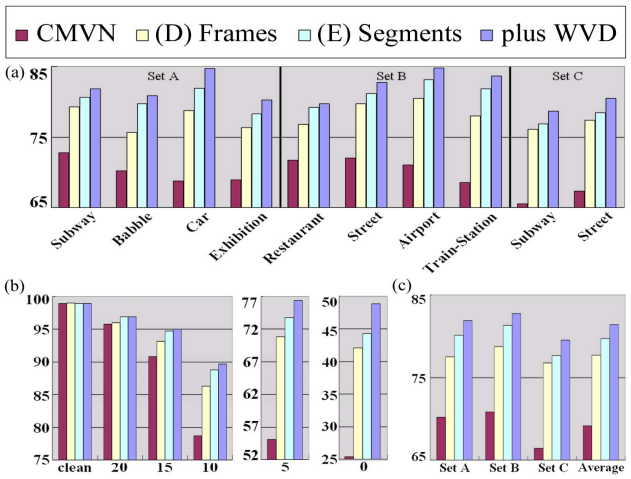

Figure 8. Incremental improvements in recognition accuracies (%) obtained with the conventional CMVN and further with the proposed approaches, (a) averaged over all SNR values but separated for different types of noise, (b) averaged over all types of noise but separated for different SNR values, and (c) separated for sets A, B, C.

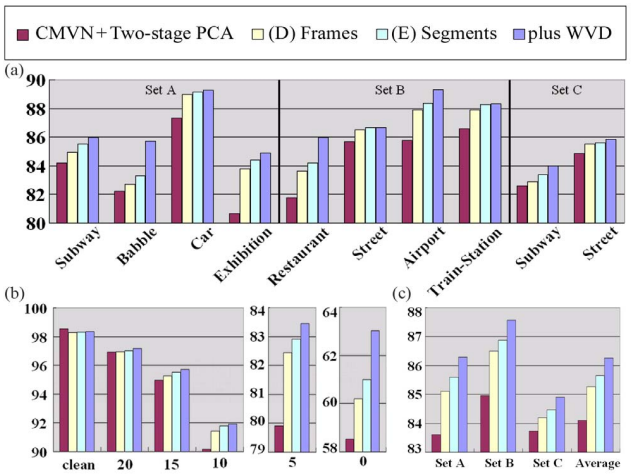

Figure 9. Incremental improvements in recognition accuracies (%) obtained with CMVN plus Two-stage PCA and with the proposed approaches applied in addition, for (a) averaged over all SNR values but separated for different types of noise, (b) averaged over all types of noise but separated for different SNR values, and (c) separated for sets A, B, C.

With weighted Viterbi decoding further applied, the most significant improvements are obtained for the babble, exhibition, restaurant, and airport cases. For example, for non-stationary airport noise the relative error rate reduction is 77.10% (from 53.25% to 89.29%) compared to the MFCC baseline result in Figure 7(a), or 24.75% compared to conventional CMVN plus Two-stage PCA in the first bar (85.77%).

In Figure 9(b), slight degradation occurred in the clean speech case, but when the SNR decreases from 20 dB all the way to 0 dB, the accuracy was improved with the proposed approaches applied one by one in addition. The effectiveness of each method becomes significant. The exact numbers for Figure 9 (b) for all SNR values are also listed in Table 1, where the last row is the error rate

reduction with respect to the results obtained with conventional CMVN + Two-stage PCA. In Table 1, the greatest improvements are obtained for the cases of 15 to 5 dB SNR, but for other SNRs the relative improvements are also significant. These results verify that the proposed approaches are useful for noisy conditions over a wide range of SNR values.

Similar observations can be made from Figure 9(c) for the three testing sets. All the above results verify that the proposed approaches can be well integrated with systems with advanced techniques.

5. CONCLUSIONS

In this paper, we propose a new approach for improved robust speech recognition by properly utilizing the reliable frames and segments obtained from noise-corrupted signals. An energy-based measure and a frame harmonicity measure are defined, and SNR-dependent GMM Classifiers are developed. We proposed various approaches to identifying reliable frames and segments, which can then be used in both front-end feature enhancement and back-end Viterbi decoding of a speech recognizer, or an advanced system with improved techniques. Very significant improvements were obtained in extensive experiments with the Aurora 2 testing environment under a wide range of noise types and SNR conditions. The results verified that the integration of these approaches can actually offer improved robust speech recognition techniques.

6. ACKNOWLEDGMENT

The authors would like to thank the reviewers for their extensive and valuable comments.

7. REFERENCES

[1] K.-T. Sung, H.-C. Wang, "A Study of Knowledge-Based Features for Obstruent Detection and Classification in Continuous Mandarin Speech," IEEE ISCSLP 2006.

[2] R. E. Yantorno, B. Y. Smolenski, A. N. Iyer, J. K. Shah, "Usable Speech Detection Using a Context Dependent Gaussian Mixture Model Classifier," IEEE ISCAS 2004.

[3] K. R. Krishnamachari, R. E. Yantorno, "Spectral Autocorrelation Ratio as a Usability Measure of Speech Segments under Co-Channel Conditions," IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), 2000.

[4] Y. Shao, D.-L. Wang, "Model-Based Sequential Organization in Cochannel Speech," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 1, pp. 289-298, Jan 2006.

[5] Y. Shao, D.-L. Wang, "Co-Channel Speaker Identification Using Usable Speech Extraction Based on Multi-Pitch Tracking," ICASSP 2003.

[6] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, no. 2, pp. 254-272, Apr. 1981.

[7] O. Viikki, K. Laurila, "Noise Robust HMM-Based Speech Recognition Using Segmental Cepstral Feature Vector Normalization," ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson, France, pp. 107-110, 1997.

[8] I. T. Jolliffe, "Principal Components Analysis," Berlin, Germany: Springer-Verlag, 1986.

[9] N.-C. Wang, J.-W. Hung, L.-S. Lee, "Data-driven Temporal Filters based on Multi-eigenvectors for Robust Features in Speech Recognition," ICASSP 2003.

SNR	20 dB	15 dB	10 dB	5 dB	0 dB	Avg
CMVN + Two-stage PCA	96.92	94.97	90.16	79.92	58.52	**84.10**
(D) Frames	96.95	95.28	91.44	82.44	60.20	**85.26**
(E) Segments	97.02	95.51	91.79	82.91	61.01	**85.65**
plus WVD	97.17	95.72	91.91	83.44	63.03	**86.25**
Total Relative Error Reduction (%)	8.01	14.91	17.79	17.55	10.87	**13.56**

Table 1. Accuracies (%) for the complete front-end in Figure 6 with the various proposed approaches applied in addition, plus the final error reduction with respect to the conventional CMVN + Two-stage PCA, for different SNR values but averaged over all types of noise.

[10] B. Raj, R. M. Stern, "Missing-Feature Approaches in Speech Recognition," IEEE Signal Processing Magazine, vol. 22, no. 5, pp. 101-116, Sep. 2005.

[11] M. P. Cooke, P. Green, L. Josifovski, A. Vizinho, "Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data," Speech Communication, vol. 34, no. 3, pp. 267-285, 2001.

[12] J. P. Barker, M. P. Cooke, D. P. W. Ellis, "Decoding Speech in the Presence of Other Sources," Speech Communication, vol. 45, no. 1, pp. 5-25, 2005.

[13] C. Cerisara, S. Demange, J.-P. Haton, "On Noise Masking for Automatic Missing Data Speech Recognition: A Survey and Discussion," Computer Speech and Language, vol. 21, no. 3, pp. 443-457, 2007.

[14] N. B. Yoma, M. Villar, "Speaker Verification in Noise Using a Stochastic Version of the Weighted Viterbi Algorithm," IEEE Transactions on Speech and Audio Processing, vol. 10, no. 3, pp. 158-166, March 2002.

[15] N. B. Yoma, I. Brito, C. Molina, "The Stochastic Weighted Viterbi Algorithm: A Frame Work to Compensate Additive Noise and Low-Bit Rate Coding Distortion," InterSpeech 2004.

[16] N. B. Yoma, F. R. McInnes, M. A. Jack, "Weighted Viterbi Algorithm and State Duration Modeling for Speech Recognition in Noise," ICASSP 1998.

[17] A. Bernard, A. Alwan, "Low-Bitrate Distributed Speech Recognition for Packet-Based and Wireless Communication," IEEE Transactions on Speech and Audio Processing, vol. 10, no. 8, pp. 570-579, Nov. 2002.

[18] X. Cui, A. Alwan, "Combining Feature Compensation and Weighted Viterbi Decoding for Noise Robust Speech Recognition with Limited Adaptation Data," ICASSP 2004.

[19] X. Cui, A. Alwan, "Noise Robust Speech Recognition Using Feature Compensation Based on Polynomial Regression of Utterance SNR," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 6, pp. 1161-1172, May 2006.

[20] Y. Chen, L.-S. Lee, "Energy-Based Frame Selection for Reliable Feature Normalization and Transformation in Robust Speech Recognition," InterSpeech 2005.

[21] A.-T. Yu, H.-C. Wang, "New Speech Harmonic Structure Measure and Its Application to Post Speech Enhancement," ICASSP 2004.

[22] S. Vaseghi, E. Zavarehei, Q. Yan, "Speech Bandwidth Extension: Extrapolations of Spectral Envelope and Harmonicity Quality of Extraction," ICASSP 2006.

[23] H.-G. Hirsch, D. Pearce, "The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," ISCA ITRW ASR2000, Paris, France, September 18-20, 2000.