TWO-DIMENSIONAL FRAME-AND-FEATURE WEIGHTED VITERBI DECODING FOR ROBUST SPEECH RECOGNITION

Yang Chang, Lin-shan Lee,

Graduate Institute of Communication Engineering, National Taiwan University
Taipei, Taiwan, Republic of China
r99942057@ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

In this paper we propose a new approach of two-dimensional frame-and-feature weighted Viterbi decoding performed at the recognizer back-end for robust speech recognition. A new SVM-based frame weighting approach is proposed considering the energy distribution and harmonicity of the frame. The feature weighting is based on a previously proposed approach using an entropy measure considering confusion between phoneme classes. These two different weighting schemes on the two different dimensions are then properly integrated in Viterbi decoding in this paper. Extensive experiments performed with the Aurora 4 testing environment showed significant improvements.

Index Terms— robust, SVM, Viterbi, weighted

1. INTRODUCTION

Robust speech recognition under noisy conditions has been an important yet unsolved problem. Many very successful feature-based and model-based approaches have been defined using more robust features or models [1-2]. Weighted Viterbi decoding [3-5] have also been proved to be useful, in which during the back-end Viterbi decoding process different weights can be assigned to the acoustic scores obtained from different frames or different feature parameters considering the discriminative power as well as the reliability of the different features or frames. A confusion-based feature weighting scheme was proposed earlier to emphasize in decoding the scores obtained with more discriminating feature parameters causing less confusion between phoneme classes [5], but this scheme didn't consider at all that some signal frames are more reliable and some others are seriously corrupted [6-7]. On the other hand, in another work different weights were assigned to different frames during decoding assuming some frames are more noisy than others [8-9], but the ways in which the weights are assigned cannot be easily learned in a new environment.

In this paper we proposed a new two-dimensional frame-and-feature weighted Viterbi decoding scheme. Reliable frames are identified and weighted higher based on an SVM classifier considering energy distribution and harmonicity of each frame. Scores obtained with more discriminating features causing less confusion between phonemes are also identified and weighted higher based on an entropy measure and a phoneme confusion matrix. Testing results on Aurora 4 showed very significant improvements. In fact, this approach is low-cost and easy to implement at the back-end decoder, therefore can be integrated with many existing feature-based and model-based approaches.

2. PROPOSED APPROACH

2.1. Overall picture

The overall picture of the proposed approach is shown in Fig. 1. The top part (Blocks (A)(B)(C)) is the Confusion-based Feature Weighting with a Training Corpus, which estimates a Confusion Matrix from the training data off-line in advance and the Confusion Matrix is used to give different weights to different feature parameters in the back-end decoding. The central part (Blocks (D)(E)) is the conventional approaches: Feature Extraction followed by Front-end Feature Normalization (e.g. CMVN or HEQ). The lower left part (Blocks (H)(I)(J)(K)) is the SVM-based Frame Weighting. It estimates the local energy distribution e_t and harmonicity h_t for each frame at time t, and trains an SVM classifier (Blocks (J)(K)) to determine a frame weighting parameter w_t for each frame. The middle-right part is then the back-end Weighted Viterbi Decoding including weighting on both feature and frame dimensions (Blocks (F)(G)). The details of these parts and blocks will be explained below.

2.2. Energy distribution and harmonicity estimation for each frame

First consider Energy Distribution Estimation in Block (H) of Fig. 1. Very often the energy distribution of a signal frame tells whether the frame is reliable or noisy. So here for each signal frame, we first calculate the smoothed instantaneous energy e[n] for each signal sample n, which is the energy averaged within a small window of length L centred on the sample being considered. In the experiment reported below, there are 200 samples in each frame and L = 11. Therefore, we have an energy vector e_t of 200 components e[n] for each frame.

We next consider Harmonicity Estimation in Block (I). The purpose is to detect the harmonic structure in the signals, which is

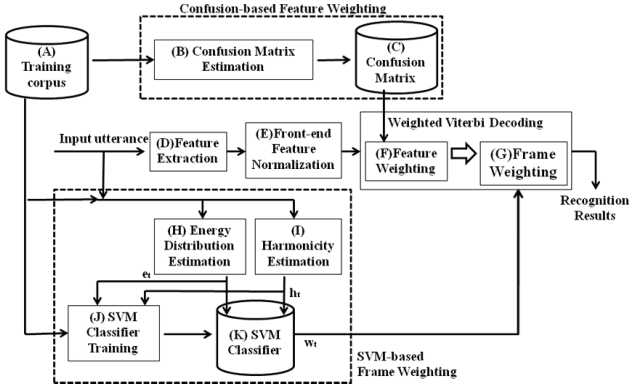

Fig. 1. Overall block diagram of the proposed approach

often a very strong indicator for voiced speech sounds, or relatively reliable parts in noise corrupted speech signals [10]. The input frames are first Hamming-windowed, low-pass filtered and transformed to the frequency domain using FFT. The squared magnitude spectrum of a frame is then cross-correlated with that of the previous frame for correlation lags ranging from -100 to 100 to give a harmonicity vector h_t with 201 components for each frame at time t. The harmonic structure of a frame may be enhanced by the cross-correlation because of the short-term stationary property of voiced speech signals [10].

2.3. Support vector machine (SVM) classifier

This is Blocks (J)(K) in Fig. 1 with a goal of giving a weight w_t to each signal frame at time t. The training of these classifiers are as follows. 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. The voiced, unvoiced speech and non-speech frames can be located on the training utterances. We then add training noise to these clean training utterances, and then calculate the energy vector e_t and harmonicity vector h_t with Blocks (H)(I) as mentioned above for each frame of these noisy utterances. The vectors (e_t and h_t) of voiced frames are taken as positive examples while those of unvoiced and non-speech frames as negative examples to train two SVMs, one with e_t and the other with h_t. In other words, we assume voiced frames are relatively more reliable than other frames in noisy speech. The weight parameter w_t for each testing frame at time t to be used in block (G) in Fig. 1. is then

$$w_t = f_1(D_e^t, D_h^t) = [exp(D_e^t)]^r \bullet [exp(D_h^t)]^{1-r}, \quad (1)$$

where D_e^t and D_h^t are the scores for the two SVMs for a testing frame t, and r is determined by a development set. So for positive frames $D^t_{e,h} > 0$ and $exp(D^t_{e,h}) > 1$, while for negative frames $D^t_{e,h} < 0$ and $exp(D^t_{e,h}) < 1$. In this way we use the SVM Classifier output to emphasize speech frames with relatively stable energy distribution and strong voicing nature, very often the nuclei of voiced phonemes, which are usually the most reliable parts in noise-corrupted speech signals.

2.4. Confusion-based feature weighting

Here we adopt the basic approaches of the confusion-based feature weighting scheme proposed earlier [5], in which scores obtained with more discriminating feature parameters causing less confusion are emphasized while others de-emphasized in the back-end Viterbi decoding, but with slight modification to the confusion matrix used. This scheme is very briefly summarized below, for lack of space [5].

We first define a confusion matrix $\{v_c(i)\}_{c,i=1}^{C}$ for each pair of monophone classes c and i, indicating how frequently each class i is misclassified as a given class c,

$$v_c(i) = \begin{cases} \log(count_c(i)+1) / \log(count_c(i_c^*)+1) & \forall i \neq c \\ 1 & i = c, \end{cases} \quad (2)$$

where $count_c(i)$ is the number of frames in the training corpus belonging to class i which are misclassified as belonging to the given class c, and C is the total number of monophone classes. The class i^*_c is the most confusing class with respect to the given class c,

$$i_c^* = \arg\max_{i \neq c} (count_C(i)). \quad (3)$$

As a result, we have $v_c(i^*_c)$ and $v_c(c) = 1$ from Eqs. (2) and (3), and $0 \le v_c(i) \le 1$ for all other i.

For each testing feature vector at frame t, $\boldsymbol{x}(t) = \{ x_d, d = 1,2,...,D \}$ (i.e., d is the parameter index and D is the total number of feature parameters), the score $p_i^{t,d}$ obtained for the d-th parameter, x_d, for a monophone class i can then be evaluated by a set of Gaussian Mixture models(GMM) for the monophone classes obtained from the clean training corpus. These scores are first normalized in Eq.(4) and then used in Eq. (5) below to evaluate a weighted entropy measure $H_c^{t,d}$ considering both the distribution $p_i^{t,d}$ for all classes i and the confusion between all classes i and a given class c,

$$\bar{p}_i^{t,d} = p_i^{t,d} / \sum_{i=1}^{C} p_i^{t,d}, i = 1,2,..., C. \quad (4)$$

$$H_c^{t,d} = -\sum_{i=1}^{C} v_c(i) \bullet \bar{p}_i^{t,d} \bullet \log(\bar{p}_i^{t,d}). \quad (5)$$

But this entropy measure $H_c^{t,d}$ in Eq. (5) is only for a given class c, therefore should be averaged over all classes c weighted by $p_c^{t,d}$ (same as $p_i^{t,d}$ are used above except i replaced by c),

$$H^{t,d} = (\sum_{c=1}^{C} p_c^{t,d} \bullet H_c^{t,d}) / \sum_{c=1}^{C} p_c^{t,d}, \quad (6)$$

and the weight parameter $W^{t,d}$ for the feature parameter x_d in the frame at time t is,

$$W^{t,d} = f_2(H^{t,d}) = \exp(-a \bullet H^{t,d}), \quad (7)$$

where a is another parameter also determined by the development set.

2.5. Two-dimensional frame-and-feature weighted Viterbi decoding

In the back-end Viterbi Decoding of Blocks (F)(G) in Fig. 1, we can now perform the two-dimensional frame-and-feature weighted Viterbi decoding to give different weights to each frame and each feature parameter. The likelihood score for a given feature vector $\boldsymbol{x}(t)$ evaluated for the j-th state of a HMM in the recognizer is

$$s[\boldsymbol{x}(t)] = w_t \bullet \log[b_j(\boldsymbol{x}(t))] = w_t \bullet \sum_{d=1}^{D} W^{t,d} \bullet \log[\sum_{m=1}^{M} c_{j,m} N(x_d(t); \mu_{jmd}, \Sigma_{jmd})], \quad (8)$$

where $b_j(\bullet)$ is the observation distribution function for the j-th state, $N(\bullet; \bullet, \bullet)$ the Gaussian component for the state and the parameter d, m the mixture index, $c_{j,m}$ the mixture weight, μ_{jmd} and Σ_{jmd} the Gaussian parameters. In Eq. (8) w_t is the frame weight in Eq. (1) and $W^{t,d}$ is the feature weight in Eq. (7).

3. EXPERIMENTAL CONDITIONS

The experiments reported here were conducted on the AURORA 4 testing environment, but only those sampled at 8 kHz. There are two sets of training data defined in Aurora 4, the clean training set and multi-condition training set, each consisting of 7138 utterances (about 12 hours). The clean training set, all recorded with the same type of microphone, was used to train the HMM models to be tested. But we created a different multi-condition training set of 7138 utterances for this research. About 25% of these utterances are clean. The rest are partitioned into 5 equal subsets and added respectively with 5 different types of additive noise: White Gaussian, pink, factory, exhibition, and train station, with SNR values uniformly distributed between 10-20 dB and an average of 15 dB. The noise types here are purposely chosen to be quite different from those observed in the Aurora 4 testing sets (car, babble, restaurant, street, airport, train station), but with similar SNR conditions (average of

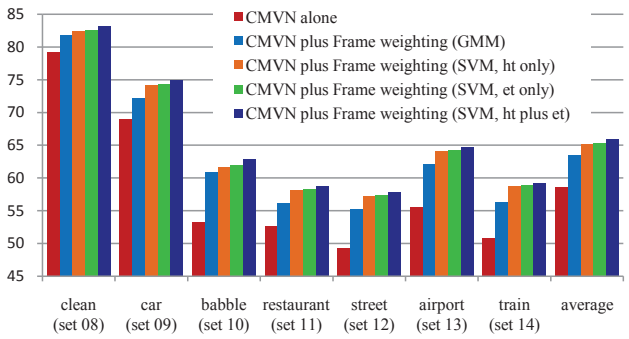

Fig. 2. Recognition accuracies (%) for Aurora 4 test sets 08-14 for frame weighting only, for CMVN alone and with different approaches applied in addition.

15 dB for Aurora 4 multi-condition training sets). This specially designed multi-condition training set was used to train the SVM classifier in Sec. 2.3 and the confusion matrix $\{v_c(i)\}$ in Sec. 2.4. This is to show that the proposed approach can offer improved recognition performance without learning the noise types in the testing set, as will be reported below. We also defined a development set to determine the parameters r in Eq. (1) and a in Eq. (7). For the development set, 100 utterances were selected from the clean training set, and then added respectively with three different types of noise (white Gaussian, pink, and factory, different from those in the test sets) with SNR ranging between 10 dB and 20dB (with an average SNR of 15 dB), thus a total of 300 noisy utterances.

4. EXPERIMENTAL RESULTS

4.1. Recognition performance with frame weighting only

We first tested the SVM-based frame weighting approach proposed here without feature weighting. The results compared with a previously proposed GMM-based approach (GMM) [9] for Aurora 4 test sets 8-14 for different noise types with mismatched microphones are in Fig. 2, in which for each noise type the first two bars are respectively for the conventional CMVN and CMVN plus the previously proposed GMM scheme [9], while the last three bars are for the currently proposed SVM-based approach (SVM), respectively using the harmonicity h_t alone, energy distribution e_t alone, and both. It can be found that the previously proposed GMM scheme offered very good improvements (bars 1 and 2), but the SVM-based approaches proposed here are clearly much better (bars 3,4,5 V.S. 2) for all types of noise. Also, for the currently proposed SVM-based approach, harmonicity h_t or energy distribution e_t alone was already better than the GMM-based scheme, while integrating the both was the best. So the two SVM classifiers respectively for h_t and e_t are complementary and additive. The average error rate reduction (last set) were 18.22% (from 58.56% to 66.11%, bars 5 V.S. 1) and 7.10% (from 63.52% to 66.11%, bars 5 V.S. 2) respectively compared to CMVN alone and the GMM-based scheme. We only show here the results for mismatched microphones for space limitation, although similar results were observed for the matched microphone (sets 1-7 of Aurora 4).

4.2. Two-dimensional frame-and-feature weighting

The results of applying the two-dimensional frame-and-feature weighted Viterbi decoding as in Eq. (8) are shown in Fig. 3. In this figure for each noise type of Aurora 4 test sets, the first three bars are for CMVN plus respectively feature weighting alone (w_t= 1 in Eq. (8)) , frame weighting alone ($W^{t,d}$ =1), and two-dimensional frame-and-feature weighting ($w_t \neq 1$ and $W^{t,d} \neq 1$). In this figure we can find that for all types of noise frame weighting was better than feature weighting, while two-dimensional frame-and-feature weighting integrating the two performed the best. So the weighting in the two different dimensions are actually additive. The average accuracy was 68.04% for the two-dimensional frame-and-feature weighting applied on top of CMVN, as compared to 64.87% for CMVN plus feature weighting alone or 66.11% for CMVN plus frame weighting alone.

4.3. Integration with other front-end feature normalization approaches

Here we replace CMVN by some other front-end feature normalization approaches, the very successful Histogram Equalization (HEQ), and the recently proposed Higher Order Cepstral Moment Normalization (HOCMN) [11]. We summarize the results in Fig. 4 again for test sets with mismatched microphones. In Fig. 4 in each set the first three bars are for CMVN, HEQ and HOCMN respectively, while the last three are for the proposed two-dimensional weighting approaches applied in addition. We see in each case the proposed approach offered very significant improvements regardless of the approach used for feature normalization. In particular, HOCMN plus the proposed two-dimensional frame-and-feature weighting always performed the best for all types of noise including clean speech (first set in Fig. 4), with an average accuracy of 70.04% (last set) or an error rate reduction of 27.70% as compared to 58.56% with CMVN alone. Note that the proposed weighting approach is simple and low-cost, easily implemented at the back-end decoder, thus can be easily integrated

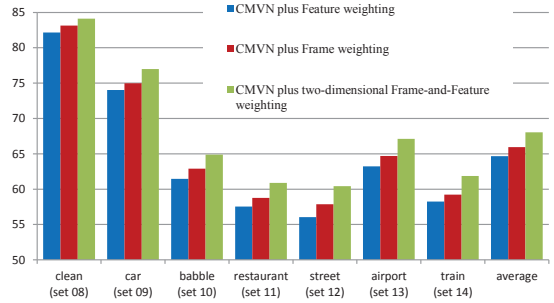

Fig. 3. Recognition accuracies (%) for Aurora 4 test sets 08-14 with feature-weighting, frame-weighting, and two-dimensional weighting, all applied on top of CMVN.

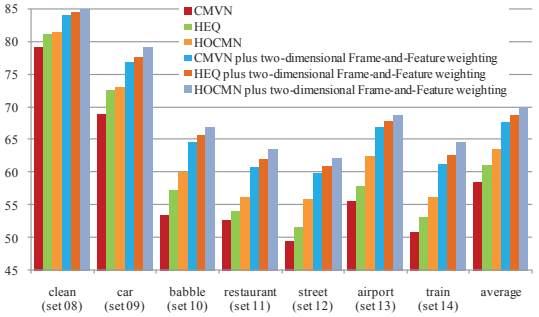

Fig. 4. Recognition accuracies (%) for Aurora 4 test sets 08-14 with CMVN, HEQ, and HOCMN, and with the proposed two-dimensional frame-and-feature weighting applied in addition.

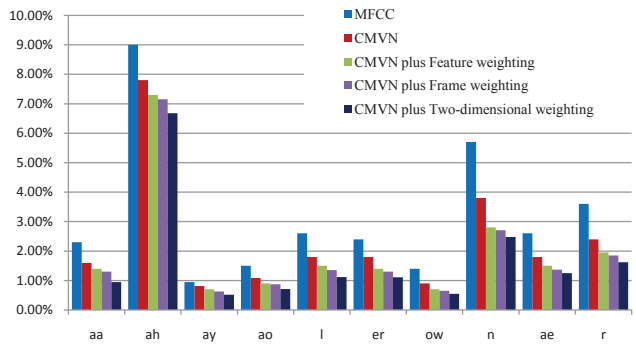

Fig. 5. Changes in numbers of misclassified frames by CMVN and the proposed approaches, normalized by the total number of misclassified frames in the MFCC baseline, separated for the 39 monophone classes, but only 10 shown here.

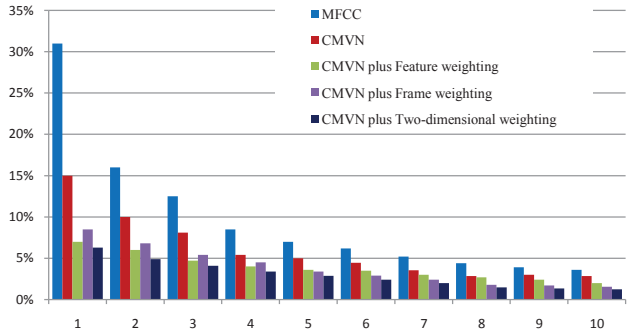

Fig. 6. Changes in numbers of misclassified frames by CMVN and the proposed approaches, evaluated for each monophone class but separated for the other 38 monophone class and sorted in order of decreasing degree of confusion, only the top 10 confusion classes are shown here.

with many very successful feature-based approaches. The results here simply verified that such integration can offer very significant improvements.

4.4. Further analysis

Fig. 5 shows the change of the number of misclassified frames by the proposed approaches for 10 out of the 39 different monophone classes. In Fig. 5, all the numbers of misclassified frames were normalized by the total number of misclassified frames for the MFCC baseline recognition task. So the first bars for MFCC baseline for all the 39 classes (10 shown here) in Fig. 5, have a sum of 100%. The second bar is for CMVN, while the last three bars for the weighted decoding applied in addition. Consistent improvements can be observed for each class shown here using individual feature or frame weighting, while the proposed two-dimensional weighting improved more. The sum of the third bars for the 39 classes is 51.36%, for the fourth bar is 49.68%, and for the last bar is 43.77%. This tells the relative reduction of misclassified frames. Consistent results can be found for the other 29 classes not shown here.

In Fig. 6, we analyze the numbers of misclassified frames separated for each monophone class but based on the degree of confusion for each class. The horizontal axis of Fig. 6 is the order of degree of confusion. For example, for the monophone class /aa/ the largest group of misclassified frames were those classified to /ah/, and the next largest group were those classified to /ay/. So these numbers contributed to the first two sets of numbers in Fig. 6 labelled as "1" and "2" on the horizontal axis. This was similarly done with all the 39 monophone classes, and these numbers from all the 39 monophone classes were added together, and then similarly normalized by the total number of misclassified frames in the MFCC baseline. Only the top 10 most seriously misclassified groups out of the 38 are shown for space limitation. From Fig. 6, we see the improvements obtained with the feature weighting (third bar) are more significant when a monophone is frequently confused with another, since the feature weighting used here considered the confusion very carefully. So the huge error reductions on the first several major error patterns (1-4 on the horizontal scale) contributed primarily to the improvements achieved by the feature weighting scheme. However, the improvements with the frame weighting scheme (fourth bar) are similar for all degree of confusion (1-10 on the horizontal scale, similarly for 11-38 not shown here). The error reductions for almost all degrees of confusion contributed to the improvements achieved by the frame weighting scheme. So weighting on the two different dimensions have different characteristics and complement each other, and integrating the two gave some further improvement (last bar).

5. CONCLUSION

In this paper, we propose a new approach for improved robust speech recognition by two-dimensional frame-and-feature weighted Viterbi decoding. An new SVM classifier was proposed for frame weighting based on energy distribution and harmonicity, while feature weighting is based on a previously proposed approach with an entropy measure considering confusion between monophone classes. Extensive experiments with the Aurora 4 testing environment under a wide range of noise types and SNR conditions showed significant improvements when applying these approaches on top of CMVN or other feature normalization schemes.

6. REFERENCES

[1] O. Viikki, K. Laurila, "Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition," Speech Communication, 1998.

[2] H. Hermansky. "Perceptual Linear Predictive (PLP) Analysis of Speech, " J. Acoust. Soc. Am. 87 (4), 1990.

[3] 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.

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

[5] Yi Chen, Lin-shan Lee, "Confusion-Based Entropy-Weighted Decoding for Robust Speech Recognition," Interspeech 2008

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

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

[8] 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.

[9] Yi Chen, Lin-shan Lee, "Robust Speech Recognition by Properly Utilizing Reliable Frames and Segments in Corrupted Signals," ASRU 2007.

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

[11] Chang-wen Hsu ,Lin-shan Lee, "Higher Order Cepstral Moment Normalization (HOCMN) for Robust Speech Recognition," ICASSP 2004.