International Conference on Spoken Language Processing, pp. 819-822, Oct. 4-8, 2004, Jeju Island, Korea

Weighting Observation Vectors for Robust Speech Recognition in Noisy Environments

Zhenyu Xiong, Thomas Fang Zheng, and Wenhu Wu

Center for Speech Technology,
State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science and Technology,
Tsinghua University, Beijing, 100084, China
[xiongzhy, fzheng]@cst.cs.tsinghua.edu.cn, wuwh@tsinghua.edu.cn

Abstract

In this paper, we propose a novel approach to robust speech recognition in noisy environments by discriminating the observation vectors. In conventional HMM-based speech recognition, all the observation vectors are treated with equal importance no matter how the corresponding speech segment is corrupted with noise. Our approach proposed here modifies the conventional decoder by weighting the likelihood scores for different observation vectors based on the signal to noise ratios (SNRs) of the corresponding speech frames when the probabilities of generating a sequence of observations are being calculated for some models. The proposed approach combined with spectral subtraction is evaluated with four different kinds of noises added to the clean speech. The experimental results show the superior performance of the proposed method over the method where only the spectral subtraction is applied, especially in the median SNR environments.

1. Introduction

One of the key issues in practical speech recognition is to improve the robustness against the mismatch between the training and testing environments [1]. The performance of speech recognition systems degrades greatly if there exists background noise, channel distortion, acoustic echo, or a variety of interfering signals. In this paper, we will focus on the environment in which the clean speech is corrupted with background noise.

Many techniques have been developed to alleviate the recognition performance degradation, such as robust feature extraction, speech enhancement, feature compensation, model compensation, and so on [2, 3, 4, 5]. The speech enhancement is the easiest way in which the noise is removed from the input noisy speech before feature extraction. Then the feature vectors extracted from the enhanced speech are decoded by the recognition models trained with clean speech, just like that the feature vectors are extracted from clean speech.

Hidden Markov Models (HMMs) and Gaussian mixture density functions are practically predominant speech recognition techniques [6]. In such systems, the probability of generating a sequence of observation vectors for some models is calculated as the product of the probabilities of generating each observation with an equal weight. In other words, each observation vector is treated with equal importance.

In noisy environments, clean speech and background noise are both time-varying. So the distortion of noisy speech is also time-varying -- speech is corrupted slightly at some time, and corrupted violently at other time. After speech enhancement, some enhanced "clean" speech is achieved. The enhanced speech is not the same as the true clean speech without distortion. While noisy speech is corrupted more violently, the enhanced speech is farther from the true clean speech. Of course observation vectors extracted form the slightly-corrupted speech should be more believable than those from the violently-corrupted speech in recognition. If the observation vectors extracted from speech with different levels of distortions can be discriminated instead of being treated with equal importance, the performance of the speech recognition system will be increased.

In this paper, we propose an approach to emphasizing the feature vectors extracted from slightly-corrupted speech by modifying conventional HMM-based decoder with likelihood scores weighted for different observation vectors. The signal to noise ratio (SNR) of the corresponding speech is used for indicating the degree of how the speech is uncorrupted.

In our speech recognition system, the input noisy speech is enhanced with the spectral subtraction (SS) method [4, 7] firstly, and SNRs for all the speech frames are estimated. Then feature vectors are extracted from the enhanced speech and decoded with acoustic models. The proposed approach is then applied to the decoding process.

This paper is organized as follows. Section 2 describes the front-end module in our speech recognition system, including speech/non-speech detection, spectral subtraction method, and SNR estimation algorithm. And Section 3 describes the proposed weighting algorithm. In section 4, the speech databases are described and the experimental results are given. Finally, in the last section conclusions are drawn.

2. Front-end Module

Figure 1 shows a block diagram of the front-end module of our recognition system. In the front-end module, a speech/non-speech detector (SND) method based on the logarithmic energy is used to classify frames as speech or non-speech firstly. Then the noisy spectrums (NS) are estimated for future use in the spectrum subtraction. And then the SNRs are estimated based on the logarithmic energies of the enhanced speech and of the estimated noise. Finally, not only MFCC-based coefficients but also the SNRs are sent to the decoder.

In the back-end module, the conventional HMM-based decoder is modified to accept SNRs to weight the likelihood scores for the observation vectors.

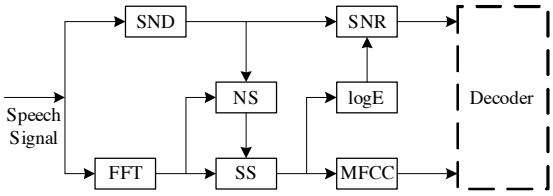

Figure 1: Block diagram of the front-end module

2.1. Quantile based speech/non-speech Detection

The detection of speech pauses is a difficult task particularly when the SNR is low. In our system, we use an approach to speech/non-speech detection derived from [8] based on the order statistics (OS) [9] filters.

Two OS filters are applied to the logarithmic energy of the signal to obtain an estimation of the local SNR of the speech signal. The first one is a median filter used to track the background noise level B. And the other one takes the *0.9*-quantile $Q(0.9)$ to track the speech level. $Q(q)$, a general expression for $Q(0.9)$, is calculated as follows. For logarithmic energies $E(t-L),\ldots,E(t+L)$ of $2L+1$ frames around the frame t to be analyzed, let $\tilde{E}(r)$, where $r=0,1,\ldots,2L$, be the corresponding values sorted in an ascending order, then $\tilde{E}(L)$ is the output of the median filter, and q-quantile is defined as

$$Q(q)=\tilde{E}(\lfloor 2qL \rfloor) \tag{1}$$

where $\lfloor x \rfloor$ denotes the greatest integer smaller than x. The difference between the speech level $Q(0.9)$ and the background noise level B is used as quantile-based estimation of the locale SNR (QSNR) of the signal.

Finally, the QSNR is compared with a threshold to make the speech/non-speech decision. If the QSNR is greater than the threshold the frame is marked as speech, otherwise as non-speech. For a non-speech frame, the background noise level B is updated using the median value obtained for this window.

And more detailed description of this speech/non-speech detection can be found in [8].

2.2. Noise estimation

Noise estimation is based on the result of speech/non-speech detection. Let $S(\omega,t)$ be the power spectrum at the frequency ω at the t-th frame of the input signal, and $N(\omega,t)$ be the power spectrum of the estimated noise at the frequency ω at the t-th frame. Only when the SND classifies the current frame as non-speech, the noise power spectrum is adapted with a forgetting factor $\lambda=0.05$ as follows

$$N(\omega,t)=\begin{cases}\lambda N(\omega,t-1)+(1-\lambda)S(\omega,t); & \text{for non-speech}\\ N(\omega,t); & \text{for speech}\end{cases} \tag{2}$$

2.3. Spectral subtraction

A traditional non-linear spectral subtraction algorithm in the power spectrum domain is used for noise reduction in the front-end as follows [7]

$$\hat{S}(\omega,t)=\max\left\{S(\omega,t)-\alpha N(\omega,t),\beta S(\omega,t)\right\} \tag{3}$$

where $\hat{S}(\omega,t)$ is the compensated power spectrum, $\alpha=1.1$ the over-subtraction factor, and $\beta=0.1$ the spectral floor.

2.4. Frame SNR estimation

The frame SNR is different from QSNR. It is based on the result of noise estimation and spectral subtraction. And it indicates the degree how the current speech frame is uncorrupted with noise. The frame SNR is defined as

$$SNR(t)=10\log\left(\frac{\sum_{\omega}\hat{S}(\omega,t)}{\sum_{\omega}N(\omega,t)}\right) \tag{4}$$

3. Weighting Algorithm

3.1. Weighting algorithm

In a conventional HMM-based speech recognition system, given a sequence of observations $\mathrm{X}=(x_1,x_2,\ldots,x_T)$ and a sequence of states $\Phi=(s_0,s_1,s_2,\ldots,s_T)$, the probability of generating the observations sequence X for the states sequence Φ is given by

$$\mathrm{P}(\mathrm{X}|\Phi)=\prod_{j=1}^{T}a_{j-1,j}b_j(x_j) \tag{5}$$

where $a_{i,j}$ is the transition probability from state s_i to state s_j while $b_j(x)$ the probability of generating observation x for state s_j [6]. In this expression, each observation is treated equally.

In order to emphasize the observations for slightly-corrupted speech, the above-mentioned expression is modified in our system as

$$\mathrm{P}(\mathrm{X}|\Phi)=\prod_{j=1}^{T}a_{j-1,j}\left[b_j(x_j)\right]^{1+\gamma_j} \tag{6}$$

where γ_j is an observation vector emphasizing weight, and its value is determined according to the SNR of the corresponding speech frame for the observation x_j. A bigger γ_j indicates a bigger importance of the observation x_j in decision.

3.2. Weighting factor

The weighting factor γ_j should be an indicator of the degree how the corresponding speech frame for the observation x_j is uncorrupted with noise. It is defined as a function for SNR of the speech frame. In our system, the speech frames of which the SNRs are higher than 20dB are considered as being less

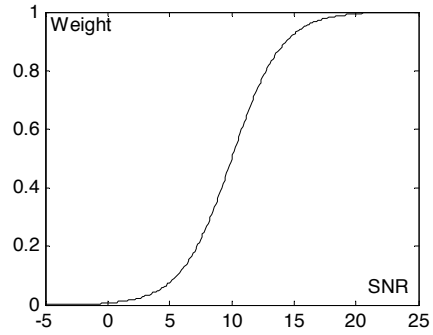

Figure 2: Weighting factor.

corrupted, and thus are emphasized greatly. On the other side, those speech frames of which the SNRs are lower than 0db are considered as being corrupted violently, and hence are not emphasized. And the other speech frames are emphasized in different levels according to their SNRs. γ_j is chosen as a sigmoid function of the SNR of x_j

$$\gamma_j = \frac{\delta}{1+\exp\left\{-\frac{SNR(x_j)-10}{2}\right\}} \tag{7}$$

where δ is a factor used to adjust the emphasizing degree, and it is experimentally chosen as 1.0 in our system. Figure 2 describes the relationship between SNR and γ.

4. Experimental Results

4.1. Speech databases and setup

Experiments are done with four different kinds of noises added to clean speech, respectively. The clean speech is an isolate words database by 20 speakers, 10 females and 10 males. Almost each speaker speaks 100 Chinese names for 4 times. The database contains 7,893 isolate word utterances in total. Four different kinds of noises, say the babble noise, the factory noise, the pink noise, and the white noise, from *Noisex* [10] are added to the clean speech with different amplitude modulation (SNR = -5, 0, 5, 10, 15, 20dB).

The experiments are carried out on a speaker-independent Chinese name recognition system. 1,429 di-IFs [11] with 3 states and a mixture of 8 Gaussian *pdf*s per state and a silence model trained with HMM Tool Kit [12] are used in this system. The acoustic model employs the 42-dimension features (containing 13 MFCC coefficients plus the logarithmic frame energy, as well as their first order and second order derivatives).

The clean and noisy data are evaluated on the conventional decoder as the baseline firstly. Then the spectral subtraction method and the proposed method are applied respectively and compared.

4.2. Experimental results and discussion

Experimental results for clean speech and four kinds of artificial noisy data are showed in Table 1 and Tables 2-5,

	Baseline
Clean	97.20

Table 1: Word accuracy (%) for clean speech.

SNR(dB)	Baseline	SS	Proposed	ERR
20	96.91	96.82	96.87	1.6
15	95.35	96.03	96.08	1.3
10	89.83	92.19	92.76	7.3
5	72.28	81.47	83.64	11.7
0	37.37	54.52	59.97	12.0
-5	12.68	23.94	26.39	3.2

Table 2: Word accuracy (%) for the Babble noisy data.

SNR(dB)	Baseline	SS	Proposed	ERR
20	96.51	97.03	97.06	1.0
15	93.71	95.36	95.68	6.9
10	85.04	91.59	92.27	8.1
5	59.43	78.79	81.39	12.3
0	24.46	49.46	55.30	11.6
-5	7.24	16.83	19.89	3.7

Table 3: Word accuracy (%) for the Factory noisy data.

SNR(dB)	Baseline	SS	Proposed	ERR
20	96.16	96.79	96.95	5.0
15	92.83	95.79	95.98	4.5
10	82.95	91.31	92.24	10.7
5	55.99	79.64	81.98	11.5
0	20.10	48.45	54.49	11.7
-5	6.1	13.92	18.02	4.8

Table 4: Word accuracy (%) for the Pink noisy data.

SNR(dB)	Baseline	SS	Proposed	ERR
20	93.28	95.02	95.19	3.4
15	87.08	92.36	92.95	7.7
10	71.19	84.76	86.2	9.4
5	39.93	66.31	69.98	10.9
0	11.59	31.98	37.60	8.3
-5	3.96	8.36	10.04	1.8

Table 5: Word accuracy (%) for the White noisy data.

respectively. In Tables 1-5, "SS" denotes the spectral subtraction method while "Proposed" the weighting algorithm combined with the spectral subtraction. "ERR" means "error reduction rate" of the proposed method compared with the spectral subtraction method.

It can be seen that the proposed method outperforms the spectral method in all kinds of experimental environments. The average relative reduction of the error rate for all experiments is about 7.1%. For higher SNR noisy environments (SNR$\geqslant$15dB), the improvement is slighter with an average ERR of 3.9%. The reason is that almost all observation vectors are slightly-corrupted and then emphasized. The emphases for different observations are counteracted. And the results are similar for low SNR noisy environments (SNR<0dB) where almost all observations are violently-corrupted and then not emphasized.

In particular, the weighting algorithm is proved effective in the median SNR noisy environments (0dB $\leqslant$SNR$\leqslant$10dB) where the average ERR for all kinds of noisy environments is

10.5%. In this case, the corrupted speech frames have different SNRs and the corresponding observations are distorted with different degrees. The proposed weighting algorithm increases the weights for the likelihood scores for the slightly-distorted observations in recognition. Then the decision is more reliable and the performance is improved.

5. Conclusions

The paper proposes an approach to robust speech recognition in noisy environments by weighting the likelihood scores for the observation vectors according to the SNRs for the corresponding speech frames. The proposed method emphasizes the importance for the observations of the slightly-corrupted speech in recognition. The experimental results show that the proposed method is superior over the one where only the spectral subtraction is applied, especially for median SNR cases (0dB $\leqslant$SNR$\leqslant$10dB).

6. References

[1] Junqua, J. C. and Haton, J. P., *Robustness in Automatic Speech Recognition*, Boston, MA: Kluwer, 1996.

[2] Acero, A., *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Boston, Kluwer Academic Publishers, 1993.

[3] Sharma, S., et al., "Feature Extraction Using Non-Linear Transformation for Robust Speech Recognition on the Aurora Database," *Int. Conf. on Acoustics, Speech and Signal Processing*, 2000, Istanbul, Turkey, pp.1117-1120.

[4] Boll, S. F., "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. on Acoustics, Speech and Signal Processing*, 1979, 27(Apr.), pp. 113-120.

[5] Gales, M.J., *Model Based Techniques for Noise Robust Speech Recognition*, PhD Thesis in Engineering Department 1995, Cambridge University.

[6] Huang. X., Acero A. and Hon, X., *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*，Prentice Hall, 2001.

[7] Lockwood P. and Boudy J., "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars", *Speech Communication*, vol. 11, pp. 215-228, 1992.

[8] Segura J. C., Benıtez M.C., Torre. A., Rubio A. J., "Feature Extraction Combining Spectral Noise Reduction and Cepstral Histogram Equalization for Robust ASR", *International Symposium of Chinese Spoken Language Processing*, pp. 225-228, 2002.

[9] H.G. Longbotham, A.C. Bovik. "Theory of Order Statistic Filters and their Relationship to Linear FIR Filters", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 37, NO. 2, February 1989.

[10] Varga, A.P., Steeneken, H. J. M., Tomlinson, M. and Jones, D., "The NOISEX-92 study on the effect of additive noise on automatic speech recognition", Technical Report, Speech Research Unit, Defense Research Agency, Malvern, UK, 1992.

[11] Xiong, Z., Zheng F., Wu. W. and Li J., "An Automatic Prompting Texts Selecting Algorithm for di-IFs Balanced Speech Corpus", *National Conference on Man-Machine Speech Communications*, pp. 252-256, Nov. 23-25, 2003.

[12] Young S., Evermann G., Kershaw D., Moore G., Odell J., Ollason D., Povey D., Valtchev V. and Woodland P., *The HTK Book*, 2002.