IMPROVING THE FRONT-END ROBUSTNESS FOR CHINESE TELEPHONE SPEECH RECOGNITION

Bowen ZHOU, Limin DU
Laboratory for Interactive Information Systems
Institute of Acoustics, Chinese Academy of Sciences
Tel: (8610)62627570 Fax: (8610)62629250 E-mail: zbw@farad.ioa.ac.cn

ABSTRACT

This paper focuses on how to improve the front-end robustness for automatic speech recognition over the telephone network. First, we study the telephone speech quality by showing SNR histogram and hum disturbance existed in telephone speech. Then we propose a simple evaluation method (CER) to assess compensation algorithm by combining the HMM and cluster analysis together in features space. At last, we discuss some improvements for classical RASTA filtering that can extract more robust features for telephone speech recognition. Based on the discussion, we advance our compensation scheme for telephone speech feature extraction.

1. INTRODUCTION

Telephone speech recognition is an increasing active research field, both for its enormous market value and for the challenges it brings to us to improve the robustness of Automatic Speech Recognition (ASR) systems. Bandwidth limitation, convolutional and additive noise, intermodulation distortion and variation of handsets and channel response all seriously degrade the performance of ASR systems. In section 2, this paper first evaluated the telephone speech quality by giving Signal to Noise Ratio (SNR) histogram of telephone speech collected via public telephone network and the disturbance of hum existed in the telephone speech. In section3, we advanced an evaluation standard based on J-measure to assess the compensation algorithm by combining the cluster analysis and HMM model. Relative Spectra (RASTA) has been adopted by many systems as front-ends for both its effectiveness and simplicity. Section 4 discusses our extended RASTA method that incorporates some improvements for classical RASTA method.

2. TELEPHONE SPEECH QUALITY EVALUATION

For the research of telephone speech recognition, we are specially establishing a Chinese telephone speech database, named as CTSIIS ([8]). Speech corpus used in this paper is all extracted from the database.

2.1 SNR histogram of telephone speech corpus

To evaluate the noise level in the public telephone network, we provide the SNR histogram of telephone speech. Figure 1 shows the global SNR histogram in a single call, which lasted for nearly 13 minutes and comprised more than 800 words that were articulated in a quite environment. It reflects the fluctuation of noise level in a single call and the global SNR histogram of all utterances from different calls in the database shows the similar result. The SNR of each utterance is calculated by measuring the average power in speech and the average power in noise and computing the ratio of average speech to noise power. To perform the average power in speech computation, the speech segments were detected first.

The histogram indicates that the range of SNR is from 17dB to 49dB. More than 90% of the utterances fall in the range of 25dB or higher SNR. As we can see in the next part, the utterances falling into the range of SNR less than 25dB were often corrupted by 50 Hz hum or static noise. In other words, additive noise is an important factor we must take into consideration for improving the robustness of telephone speech recognition.

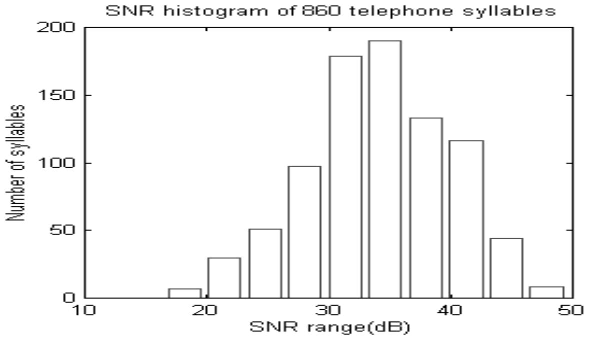

Figure 1. The SNR histogram of single call

2.2 50Hz hum interference

In the Mainland China, the basic frequency of alternating current is 50 Hertz. The interference by hum of 50 Hertz or its harmonic is a major source of degradation of speech recognition in the public telephone network. We analyzed the speech corpus collected from different calls. For each call, we identify the speech and non-speech segments. Then we divide the non-speech segments into 100 ms windows and compute the normalized auto-correlation sequence (via FFT) in each window. Finally we average all these sequences together and look for peaks at lag values corresponding to frequencies of interest: 50 Hertz and its harmonic. Referring to the results presented in Figure 2, we can see that there are pointed pulse at the frequency of 50 Hertz and relative high peaks at frequency of 250 and 300 Hertz. We can conclude that sometimes these hum artifacts seriously degrade the telephone speech and should be checked out.

3. EVALUATING COMPENSATION ALGORITHM PERFORMANCE USING J-MEASURE

For telephone speech recognition system, the channel effects increase the overall variability of the feature vectors which must be handled by the HMM. In other words, channel effects impair the discriminating ability of feature vectors.

It is postulated that high correlation exists between feature sets that have good measured class discrimination and those that give good recognition accuracy. J-measure ([5]) is a kind of criterion to evaluate the discrimination of a given feature set. In this section, we propose a method to evaluate the compensation effect of front-end compensation algorithm by comparing the J-measures of features compensated by the algorithm and original features (For instance, RASTA-PLP and PLP).

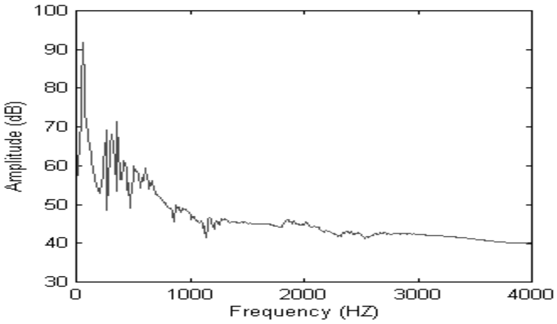

Figure 2. Energy distribution of silent parts of telephone speech

We selected an evaluating database comprising 75 isolated syllables from CTSIIS. Hidden Markov Models (HMM) were trained using the feature of PLP or RASTA-PLP respectively to represent each word of the database. Each state of each model is treated as a separate speech class. With these models, for testing speech, we can decode the feature frames and make each frame corresponding to one state. Feature frames that belong to the identical state will be treated as the same class. Then we can calculate Matrix B, the between-class covariance for each word and Matrix W, the within-class covariance for PLP and RASTA-PLP respectively. At last, J-measure, defined as $J = tr(W^{-1}B)$, can be calculated individually as $J_{plp}, J_{rasta-plp}$.

We define the Compensation Effectiveness Ratio (CER) as:

$$CER = \sum_{i=1}^{N} (J_{rasta-plp} / J_{plp})_i \qquad (1)$$

Where, N is the size of evaluating database, equaling

75 here. Obviously, the more the CER, the more effective the compensation algorithm.

4. EXTRACTING MORE ROBUST FEATURES

RASTA ([1]) is an effective method to suppress the convolutional distortion introduced by telephone channel and therefore it has been adopted by many systems as front-ends. However, there is still room for improvement. We can achieve more robust features for telephone speech recognition by extending RASTA through the following ways.

- **Preprocessing for telephone speech**

As we can see in section 2, telephone speech suffers from the disturbance of hum and ambient additive noise. Preprocessing such as comb filtering to suppress the hum of 50 Hertz will be helpful. For additive noise, RASTA (J-RASTA) processing tries to address the convolutional noise and additive noise at the same time. However, since J-RASTA is a trade-off between two kinds of effects, it can not compensate for both of them completely. In other words, it could improve the robustness of RASTA to compensate the spectrum before RASTA in linear or logarithmic domain. For instances, spectral subtraction (SS) is widely used for additive noise suppression. Because of its simplicity, SS is easy to use for noise suppression and it can work well as preprocessing with another feature extraction technique such as RASTA.

The scheme of spectral subtraction we use is defined as follows:

$$Y_{ss}(\omega,i) = \max(Y(\omega,i) - \alpha\tilde{N}(\omega,i), \beta Y(\omega,i)) \quad (2)$$

Where, α is an over-estimated factor, β is a flooring factor, $\tilde{N}(\omega,i)$ is smoothed noise estimation of current speech segment and:

$$\tilde{N}(\omega,i) = \gamma N(\omega,i-1) + (1-\gamma)N(\omega,i)$$

Typical values of α, β, γ can be set as 1.0, 0.1, and 0.2 respectively.

- **Phase correction for RASTA filtering**

It is very important to preserve the phase information in modulation frequency domain ([4]) while the phase response of classical RASTA filter is not flat in the region of important modulation frequencies between 1Hz and 16Hz. The all-pass phase correction filter ([3]) followed classical RASTA filter maintains the original magnitude response and flats the phase response, so it can preserve original phase information as much as possible. For standard RASTA filter in equation (3),

$$H(z) = 0.1z^4 * \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.94z^{-1}} \quad (3)$$

The all-pass phase correction filter can be implemented as a pole-zero filter:

$$H_{pc}(z) = \frac{b_0 + b_1 z^{-1} + \cdots + b_q z^{-q}}{a_0 + a_1 z^{-1} + \ldots + a_p z^{-p}} \quad (4)$$

For q=1 and p=7, we can calculate the coefficients as:

Coefficients	B	A
0	-17.7824	1.0000
1	17.3362	9.2453
2	—	-4.5402
3	—	18.1181
4	—	-6.7073
5	—	2.6271
6	—	0.7388
7	—	0.0769

In this manner the left-context dependency introduced by RASTA filter is also removed. Experimental results suggest that phase corrected RASTA can outperform classical RASTA significantly ([3]).

- **Linear Discriminant Analysis (LDA)**

LDA ([5], [7]) is a data driven orthogonalization process designed to maximally discriminate classes by means of linear transformation. This method maximizes the ratio of between class variance to the within class variance in any particular data set thereby guaranteeing maximal separability. The long integration window of the RASTA filter may have lead to wider class clusters with more possibility of class overlap. Consequently, combination of RASTA and LDA should improve the ability of features to discriminate. The process of LDA contains three principal steps (see in [5]) and it can be computed using separate programs that operated on the RASTA feature files.

These improvements we discussed above are proposed for improving the robustness of RASTA features from different perspective. It maybe implies that combination of all these extensions will cause more advancement than any single one. Therefore, we propose our scheme for extracting telephone speech features, which was presented in figure 3. It was still need to be verified. The evaluation for this scheme is in processing now, by comparing it with classical RASTA method for both the recognition results and CER value.

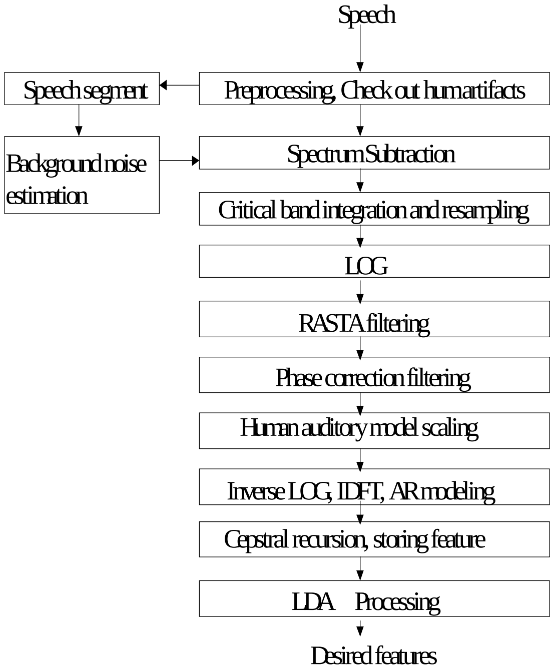
Speech
Speech segment
Preprocessing, Check out hum artifacts
Background noise estimation
Spectrum Subtraction
Critical band integration and resampling
LOG
RASTA filtering
Phase correction filtering
Human auditory model scaling
Inverse LOG, IDFT, AR modeling
Cepstral recursion, storing feature
LDA Processing
Desired features

Figure 3. The main processing steps for extended RASTA method

5. CONCLUSION

This paper studies the telephone speech quality first with the aim of improving the recognition performance over the telephone network. We concluded that both additive noise and hum of 50 Hertz degrade the speech quality. Then we proposed a simple evaluation method (CER) to assess compensation algorithm by combining the HMM and cluster analysis together in features space. At last, we proposed our scheme for extracting more robust features by extending classical RASTA method.

6. ACKNOWLEDGEMENT

The authors would like giving the most sincere thanks to Dr. Ying JIA for the revelatory discussions of HMM and cluster analysis with us. Thanks also to Dr. Johan de Veth for the discussions of phase-corrected RASTA.

7. REFERENCE

[1.] H. Hermansky and N. Morgan, “RASTA processing of speech”, IEEE Trans. On Speech and Audio Processing, Vol. 2(4), pp. 578-589,1994

[2.] S. Nicholson, et al, “Evaluation feature set performance using the F-ration and J-measures”, In Proc. of Eurospeech, 1997

[3.] Johan de Veth et al, “Phase-corrected RASTA for automatic speech recognition over the phone”, In Proc. ICASSP, 1997

[4.] Noboru Kanedera, et al, “ On properties of modulation spectrum for robust automatic speech recognition”, In Proc. ICASSP, 1998

[5.] T. W. Parson, “ Voice and speech processing”, McGraw-Hill Book Company, New York, 1986

[6.] Hiroaki Ogawa, “ More robust J-RASTA processing using spectral subtraction and harmonic sieving”, Technical Report TR-97-031, ICSI, Berkeley

[7.] Michael L. Shire, “Deployment of RASTA-PLP with the Siemens ZT speech recognition system”, Technical Report TR-97-057, ICSI, Berkeley

［8.］ Bowen Zhou and Limin Du, “ CTSIIS: A Chinese telephone speech collecting system and corpus”**,** Proceedings of the Fifth National Conference on Man-Machine Speech Communication, Haerbin, China, 1998