RECONSTRUCTING MISSING SPEECH SPECTRAL COMPONENTS USING BOTH TEMPORAL AND STATISTICAL CORRELATIONS

Mohamad Mohsen Goodarzi, Farshad Almasganj, Mohammad Ahadi

1Amirkabir University, Department of Biomedical Eng.  
2Amirkabir University, Department of Electrical Eng.

ABSTRACT

This paper presents a new method for reconstructing unreliable spectral component which uses statistical distributions of former and later reliable frames and reliable components of current frame. In this technique, first, a HMM is used to model the temporal variation of clean speech signal. Then using this model and according to probabilities of occurring noisy component at each states, a distribution for noisy components is estimated. Finally, by applying MAP estimation on mentioned distribution final estimation of this unreliable component is obtained. The proposed method has been compared to a recent missing feature method which is based on clustering feature vectors and exhibits a significant enhancement in two different noisy environments.

1. INTRODUCTION

The performance of automatic speech recognition (ASR) systems is adversely affected by noise and a wide range of different algorithms has been proposed to cope with the effects of additive noise. While these approaches are reasonably effective in the context of their intended purposes, they are generally ineffective in many more difficult environments. Missing feature is one of recent approaches which has been concerned during last year's. These approaches are based on omitting the most noise affected components of speech spectral and then reconstructing them using the other reliable components. Investigations have shown that human listeners are able to comprehend speech that has undergone considerable spectral excitations. For example, normal conversation is possible with speech that has been either high- or low-pass filtered with a cutoff frequency of 1,800Hz [1].

Being based on redundancy of speech signal rather than characteristics of noise makes these approaches independent of the type of additive noise.

Missing feature approach has been proposed at the University of Sheffield in the mid-1990s. In the two original algorithm proposed by Sheffield group, recognition was performed by HMM-based recognizers directly with the incomplete spectrographic information from reliable time-frequency components. Since conventional HMM-based recognizers cannot perform recognition with incomplete representations, their algorithm modified the manner in which state output probabilities were computed within the recognizer [2].

Cooke et al use the value of any unreliable time-frequency components as an upper bound of that component based on this assumption that noise is additive and uncorrelated to speech signal [3]. These two algorithms and similar ones which modify recognizer models are called model-based missing feature methods. But suffering of drawbacks such as incompatibility with MFCC and need to modifying recognizer model, made these methods less appealing.

Raj et al proposed two methods [4] to reconstructing incomplete spectrogram without any necessary change in recognizer model. Also reconstructed features can be transformed to MFCC or any other arbitrary features. These algorithms have been known as feature-based missing feature.

Rodbro et al utilized the sequential variation of neighboring frames to estimate and reconstruct the lost packet data of VoIP using HMM [5].

Using a similar approach, in [6] HMM has been used to reconstruct noisy components of speech spectral. But during estimating the unreliable components of any feature vector at any specific state, the mean of that state has been used as estimation of that frame which means ignoring reliable components of that feature vector and replacing all components (either reliable or unreliable) with mean vector of current state.

The proposed technique in this paper uses HMM to obtain a distribution of noisy component which is mixed of all states distribution and then estimating that component by applying MAP on yielded distribution. This technique has been compared to proposed method at [4] which is one of most powerful one at the missing feature methods and is known as cluster-based.

Section 2 introduces the details of proposed technique and the cluster-based technique.
The results of applying each method at two different environments (white noise and babble noise) at five different SNR are presented in section 3 and in section 4 concluding remarks are given.

2. RECONSTRUCTING UNRELIABLE COMPONENTS

2.1 Cluster-based reconstruction

As discussed in previous section after Sheffield group researches, wide variety of missing feature methods has been proposed. Cluster-based is one of its recent and powerful ones. In this method first, clean data is clustered using a clustering method such as K-means. Then statistical characteristics of each cluster should be calculated. Now we can estimate unreliable components of noisy feature vector according to equation (1).

\[
\hat{X}_u^k(t) = \arg \max_{X_u} P(X_u(t), X_u(t) \leq Y_u(t) | k, Y_r(t))
\]

(1)

Where \( Y \) is observed spectral vector, \( X \) is corresponding clean vector, subscript \( k \) represents \( k \)-th cluster and subscripts \( u \) and \( r \) represent unreliable and reliable component respectively.

Equation (1) is a bounded MAP estimation and is calculated according to proposed algorithm in [7].

Eventually, final estimation of missing components is calculated through following equation.

\[
\hat{X}_u(t) = \sum_{k=1}^{K} P(k | Y_r(t), X_u(t) \leq Y_u(t)) \hat{X}_u^k(t)
\]

(2)

In equation (2), term \( P(k | Y_r(t), X_u(t) \leq Y_u(t)) \) used as weight for obtained estimation of cluster \( k \) and calculated using Bayes rule according to [4].

2.2 Proposed method

Although cluster-based reconstruction is a powerful technique but also suffer from an important defect which is lack of using temporal correlation between neighboring components.

In proposed method, clean speech data is modeled using a k-state HMM. So we have initial state distribution \( \pi_i \), transition probability \( a_{ij} \) and the observation probability distribution \( b_i(Y') \). \( i \) and \( j \) represents the state number. Now we segment the frame sequence of observed utterance to parts with length of \( T \) frames and assume that each part has been generated by mentioned HMM, so forward and backward parameter for each part can be computed as [8]:

\[
\alpha_{t+1}(j) = \left[ \sum_{i=1}^{K} \alpha_i(i)a_{ij} \right] b_j(Y'(t+1))
\]

(3)

\[
\beta_t(i) = \sum_{j=1}^{K} a_{ij}b_j(Y'(t+1))\beta_{t+1}(j)
\]

(4)

But one can say how above parameters can be obtained while some of components of \( Y(t) \) are unreliable. To do this, we suggested to compute a MAP estimation of \( Y(t) \) according to the distribution of \( i \)-th state which yields \( \hat{Y}_{MAP}^i(t) \) and then we can compute \( \beta_t(i) \) and use it instead of \( \beta_t(i) \) in equations (3) and (4).

Now assume that we have an utterance with the following frame sequence: \( Y(1), Y(2),..., Y(t),..., Y(T) \), and we want to reconstruct unreliable components of \( Y(t) \). The probability of being at \( i \)-th state at \( t \)-th moment and observing above sequence would be:

\[
\gamma_t(i) = \frac{\alpha_i(i)\beta_t(i)}{\sum_{i=1}^{K} \alpha_i(i)\beta_t(i)}
\]

(5)

On the other hand, the MAP estimation of unreliable component of \( Y(t) \) according to distribution of \( i \)-th state is \( \hat{Y}_{MAP}^i(t) \). Thus averaging over all possible condition would yield the final estimation as:

\[
\hat{Y}(t) = \sum_{i=1}^{K} \gamma_t(i) \hat{Y}_{MAP}^i(t)
\]

(6)

Where \( \alpha_i(i) \) and \( \beta_t(i) \) are forward and backward parameter at \( t \)-th frame and \( i \)-th state respectively.

3. EXPERIMENTAL RESULTS

Methods of section 2 have been applied on FarsDat database which contains utterances from 304 different speakers each ones uttered 20 sentences.

Features are extracted at two sets. First set contains 13 MFCC coefficients plus 13 delta plus 13 accelerations coefficients. This features set has been used to train recognition system. Applying cepstrum on features can disturb the correlation between them and as discussed at pervious section we need this correlation as an important factor for reconstructing so we define second set of features as mel-log-spectrum which is obtained by applying mel filterbank on speech spectrogram. This set has been used as train and test data for reconstructing system and after being reconstructed is transformed to the type of first set to be recognized.

We used HTK as recognition system and trained it using 5000 sentences of database. Recognition has been done at word level. White and babble noise from NOISEX is utilized to prepare test data.

Performance of two mentioned techniques in white noise polluted speech at 5 different SNRs is depicted in
Figure 1. Baseline indicates the result of recognition without applying any reconstructing method.

Figure 1. Comparing recognition accuracy for white noise and with real local SNRs.

Figure 2 shows their performance in condition of babble noise. Finding unreliable components of observed spectral vectors is a challenging problem in missing feature methods. At the results of figure 1 and 2 we used an SNR threshold to classifying components as reliable or unreliable and SNR has been obtained using clean speech.

Figure 2. Comparing recognition accuracy for babble noise and with real local SNRs.

There are many divers of proposed methods to distinguish reliable and unreliable components [9]. Here we used a simple one to apply a more practical test on our proposed method. To do this we assume first 20 frames of each utterance as silence and their mean as estimation for noise. So we can obtain local SNRs and use those to distinguish between reliable and unreliable components. Figure 3 and 4 shows the result based on these local SNRs.

Figure 3. Comparing recognition accuracy for white noise and with estimated local SNRs.

Figure 4. Comparing recognition accuracy for babble noise and with estimated local SNRs.

4. CONCLUSIONS

This paper proposed a technique to reconstructing unreliable spectral components of a speech signal and to do that we take advantages of both statistical correlation between components in a frame and the temporal correlation between adjacent frames. Figures 1 and 2 illustrate the effectiveness of proposed technique in two different environment (e.g. White and babble noise). The proposed technique has a significant preference compared to cluster-based method in all SNR levels. Also comparing figures 1 and 2 to figures 3 and 4 illustrates the robustness of this technique related to type of noise.

Noticing to figure 1, we see that performance of cluster-based method decreases at high SNRs while proposed technique has a good performance at high SNRs and won't decreases the performance of recognition system for clean data.

In the results of figures 3 and 4, we used a simple method to estimate local SNRs which caused to decrease the results compared to figures 1 and 2. But the point should be considered here is the partial independency of proposed technique to the method of finding unreliable components.
5. REFERENCES


