On the Mask Modeling and Feature Representation in the Missing-Feature ASR: Evaluation on the Consonant Challenge

Peter Jančovič, Műinever Köküer

Electronic, Electrical & Computer Engineering, University of Birmingham, Birmingham, UK

\{p.jancovic, m.kokuer\}@bham.ac.uk

Abstract

In this paper we investigate an incorporation of mask modeling within the missing-feature HMM-based ASR and also explore on feature representation within this framework. The mask model is estimated for each HMM state and mixture by using a separate Viterbi-style training procedure. We explore an employment of the frequency-filtered features and their combination with the logarithm filter-bank energies. Experimental evaluation is performed on the Consonant Challenge corpus. The obtained results show significant improvements by incorporation of the proposed methods over the standard MFT-based ASR.

Index Terms: automatic speech recognition, missing-feature theory, mask modeling, mask-probability, frequency-filtering, consonant identification, VCV, noise-robustness.

1. Introduction

The performance of automatic speech recognition (ASR) systems degrades rapidly when speech signal is corrupted by a background acoustical noise. Traditionally, robustness of ASR systems to noise can be improved by employing speech enhancement techniques to remove the noise, for example, spectral subtraction, Wiener filtering or MAP-based enhancement, or noise-compensation techniques to compensate for the effect of the noise in feature or model domain.

Recently, the missing-feature theory (MFT) has been proposed for dealing with noise corruption in speech and recognition, e.g. \cite{1} \cite{2}. The MFT assumes that the feature vector characterizing the signal contains elements which are not affected by noise, referred to as present, and elements corrupted by noise, referred to as missing. This information is contained in the so-called mask. The effect of noise may be suppressed by marginalizing or imputing the elements assigned as missing in the mask in the observation probability calculation. The performance of the MFT method depends critically on the mask accuracy. The mask may be obtained in various ways, for instance, using an estimate of the local-SNR based on a noise estimate \cite{2} \cite{3}, an estimate of inherent properties of speech signal such as harmonicity \cite{4} \cite{5} or employing a separate mask classifier \cite{6}. The computational auditory scene analysis may also be employed in the process of mask estimation.

Another important issue in the MFT is the feature representation. As the employment of the MFT requires the noise-corruption to be localized into several features, typically, the log filter-bank energies (logFBEs) have been employed within the MFT framework. However, they suffer from a high correlation between the features, which makes the diagonal covariance matrix modeling not appropriate. Some studies have recently been presented on the use of MFT in the cepstral domain \cite{7} \cite{8}. Recently, an employment of the frequency-filtering (FF) technique, introduced in \cite{9}, has been employed in the MFT framework, e.g. \cite{10}.

In this paper, we investigate two issues in the MFT-based ASR, mask modeling and feature representation, and present experimental evaluations on the Consonant Challenge. First, we investigate an incorporation of a mask model within the MFT framework. This has not been considered in previous works on the MFT. The mask model, estimated for each HMM state and mixture, expresses what mask the state would expect in a given noisy condition. Second, we study the feature representation employed in the MFT-based ASR. First the FF features and then a combination of FF features with logFBEs are explored. The combined feature representation can capture both the slope of the spectral envelope and information about energy level. Experimental evaluations are performed on the Consonant Challenge recognition by employing the marginalisation-based MFT. To demonstrate the performance of the proposed methods, evaluations on all testsets are performed using an oracle mask and on testset 4 also using an estimated mask based on a simple onset detection. It is demonstrated that both of the investigated methods, mask modeling and feature representation, can provide significant performance improvement over the standard MFT-based ASR. Using the oracle mask, the performance is in many cases relatively close to the human consonant recognition accuracy.

2. Automatic speech recognition based on missing-feature theory

Let us consider $Y$ to be the sequence of observation vectors extracted from a given speech utterance. The goal of a speech recognizer is to find the word sequence $\hat{W}$ that maximizes the posterior probability $P(W|Y)$. Assume that some features in $Y$ are missing and this information is contained in the mask $M$. Considering an HMM-based ASR system employing the missing-feature theory (MFT), the search for $\hat{W}$ can be expressed as

$$\hat{W} \approx \arg \max_{W} P(Y|M, S, W) P(M|S, W) P(S|W) P(W)$$

(1)

where $P(S|W)$ is the HMM state-transition probability, $S$ is the sequence of HMM states used during the recognition, and $P(W)$ is the language-model probability. The term $P(Y|M, S, W)$ is the probability of the incomplete observation sequence $Y$ which may be calculated as summarized in the following subsection. The term $P(M|S, W)$ is referred to as mask-probability and it expresses how likely the given mask $M$, corresponding to the observation sequence $Y$, is being generated by the HMM state sequence $S$. This term is one of the
issues studied in this paper and is presented in the following section.

2.1. Marginalisation-based missing-feature model

Let us consider that speech is modeled by an HMM with state output probabilities modeled by a mixture of Gaussian distributions with diagonal covariance matrices. Let \( y_t \) be the feature vector at frame-time \( t \) from the sequence \( Y \), and \( m_b \) be the corresponding mask vector from the sequence \( M \) determining whether an element of \( y_t \) belongs to the speech signal, i.e., present, or belongs to noise signal, i.e., missing.

In the marginalisation-based MFT, the likelihood \( P(y_t|s, l, m) \) of the feature vector \( y_t \) at state \( s \) and mixture component \( l \) is calculated by integrating out the missing elements of the feature vector \( y_t \), i.e.,

\[
P(y_t|s, l, m) = \prod_{b \in \text{prec}} P(y_t(b)|s, l)
\]  

Considering spectral energy features, the actual feature value can be considered as an upper-bound and zero as a lower-bound in the integration of the missing features [12]. This is referred to as bounded marginalisation. The likelihood \( P(y_t|s, l, m) \) can then be calculated as

\[
P(y_t|s, l, m) = \prod_{b \in \text{prec}} P(y_t(b)|s, l) \prod_{b \in \text{miss}} \int_0^{y_t(b)} P(y_t(b)|s, l) \, dB
\]  

2.2. Mask estimation for filter-bank channels

The mask \( M \) required by the MFT model may be obtained in various ways as summarised in Section 1. As the mask estimation is not of the main focus of this paper, we employed the oracle mask and a mask estimated based on a simple onset detection to demonstrate the potential of the methods studied.

The oracle mask is derived based on full a-priori knowledge of the noise and clean speech signal and as such indicates an upper-bound performance. We used the a-priori SNR to construct the oracle mask as

\[
m_t(b) = 1 \quad \text{if} \quad 10 \log \left( \frac{X_t(b)}{N_t(b)} \right) > \gamma.
\]

where \( X_t(b) \) and \( N_t(b) \) are the filter-bank energy of the clean speech and noise, respectively. Threshold \( \gamma \) was set to -6dB as it was shown to provide a good performance in our experiments.

The onset mask was estimated based on a simple comparison of the filter-bank energy of the noisy signal, \( Y_t(b) \), with a threshold determined based on the median energy of the channel, \( E(b) \), and median of the five highest energies, \( \bar{E}(b) \), over time as

\[
m_t(b) = 1 \quad \text{if} \quad Y_t(b) > \gamma(b)
\]

where the threshold \( \gamma(b) \) was set to 0.5(\( \bar{E}(b) - \bar{E}(b) \)) + \( \bar{E}(b) \).

3. Incorporation of mask modeling within the missing-feature ASR

This section presents the proposed incorporation of the mask-probability term \( P(M|S, W) \) in the MFT-based HMM framework.

At a given frame-time \( t \), the mask-probability \( P(m_b|s, W) \) expresses the probability that the mask \( m_b \) is generated by the HMM state \( s \). As such, the incorporation of the mask-probability can be beneficial in situations when an incorrect model would accidentally produce high likelihood for the given features, however, its mask model would not agree with the mask associated with these features.

Having an example of noise (or knowledge of noise characteristics), the mask model could be estimated based on masks obtained from the training data corrupted by the given noise. Having no information about noise, it could be estimated by using a mask reflecting some a-priori knowledge about speech, for instance, the fact that high-energy regions of speech spectra are less likely to be corrupted by noise. In this paper, we performed multi-condition training of the mask model in experiments with oracle masks and matched-condition training in the case of onset masks. Noise signals, extracted from all noisy conditions of the development set, were used to contaminate the training set; the oracle masks obtained based on these corrupted training data were then used for the mask-probability estimation.

The estimation of the mask-probability is performed by a separate training procedure that is performed after the HMMs have been trained (i.e., the trained HMMs are not altered). The following sections give detailed description of the estimation of the mask-probability and its incorporation during the recognition.

3.1. Estimating the mask-probability for HMM states

The estimation of the mask-probability \( P(m|s, l) \) at each HMM state and mixture can be performed using the noisy training data, as discussed above, by a Baum-Welch or Viterbi-style training procedure; the latter was used in this paper.

Given a speech utterance, we have a sequence of feature vectors \( \{y_1, \ldots, y_T\} \) and the corresponding sequence of mask vectors \( \{m_1, \ldots, m_T\} \) where \( T \) is the number of frames. The Viterbi algorithm is then used to obtain the state-time alignment of the sequence of feature vectors on the HMMs corresponding to the speech utterance. This provides an association of each feature vector \( y_t \) to some HMM state \( s \). The posterior probability that the mixture-component \( l \) (at the state \( s \)) has generated the feature vector \( y_t \) is then calculated as

\[
P(l|\gamma_t, s) = \frac{P(y_t|s, l)}{P(l|s)} = \frac{P(y_t|s, l)}{\sum_{t'} P(y_t|s, t') P(l|s)}
\]  

where the mixture-weight \( P(l|s) \) and the probability density function of the features used to calculate the \( P(y_t|s, l) \), are obtained as an outcome of the HMM training.

For each mixture \( l \) and HMM state \( s \), we collect (over the entire training dataset) the posterior probabilities \( P(l|y_t, s) \) for all \( y_t \)’s associated with the state \( s \) together with the corresponding mask vectors \( m_t \)’s. The mask-probability of the \( b^{th} \) feature can then be obtained as

\[
P(m_b = a|s, l) = \frac{\sum_{y_t \in s} P(l|y_t, s) \delta(m_b(a), a)}{\sum_{y_t \in s} P(l|y_t, s)}
\]  

where \( a \in \{0, 1\} \) is the value of mask and \( \delta(m_b(a), a)=1 \) when \( m_b(a)=a \), otherwise zero.

3.2. Mask-probability incorporation during recognition

The value of the mask-probability when being incorporated in the overall probability calculation in Eq. 1 may need to be scaled in order to achieve an appropriate effect of the mask-probability on the overall probability (akin to language model scaling). This can be performed by employing a sigmoid function to transform the \( P(m_b|s, l) \) for each \( b \) to a new value, i.e.,
where $\alpha$ is a constant defining the slope of the function and the value 0.5 gives shift of the function. The bigger the value of $\alpha$ is the greater the effect of the mask-probability on the overall probability. An appropriate value for $\alpha$ can be decided based on a small set of experiments on a development data; $\alpha=4$ was used for all the experiments reported in this paper.

4. Feature representations for the missing-feature ASR

The employment of the MFT requires the noise-corruption be localised into several features. Due to this, often the log filter-bank energies (logFBEs) have been employed within the MFT framework, despite of being highly correlated. In this paper, the employment of the frequency-filtered (FF) features and their combination with the logFBE features, as discussed below, are studied in the MFT-based ASR.

4.1. Frequency-filtered features

The frequency-filtering technique [9] [13] performs filtering of the logFBEs, with an effect of decorrelating them. Various types of frequency-filter may be used. It has been reported in [9] [13] that the FIR filter with a transfer function $H(z) = z^{-1}$ produces good results across many different tasks and databases. This filter has been employed to generate the FF features in our experiments. Given a $Q$-dimensional logFBE feature vector, the FF feature vector is of $Q - 2$ dimension as the edge values of the filtering operation have been discarded.

4.2. Combination of the FF and logFBE features

The FF features capture the information about the slope of the spectral envelope. However, they lose the information about whether the slope appears at high or low energy regions. As such, models having similar slope may obtain high likelihood although the corresponding regions may have been of completely different energy level. To overcome this problem we investigate the integration of the energy level information by appending the logFBEs into the FF feature vector.

Note that authors in [14] have reported that the appending of the logFBEs to the FF features decreased the recognition performance in standard HMM-based ASR system. However, this may have been due to the fact that the incorporation has not been performed within the MFT framework as is investigated in this paper.

Note also that the employment of the combined FF and logFBE features within the MFT framework allows us to perform bounded marginalisation on the logFBE features. These experiments have been conducted and results are reported in the experimental section.

5. Experimental evaluations

5.1. The Consonant Challenge corpus

Experiments were performed using the corpus from the Inter-speech 08 Consonant Challenge [15]. The Consonant Challenge corpus consists of intervocalic English consonants (VCV), for a number of vowel and stress combinations. The 24 consonants combines with nine vowel contexts consisting of all possible combinations of the three vowels.

5.2. Experimental set-up

Experiments were performed by employing two types of feature representation: the FF features and the combined FF+logFBE features, each being appended with the first and second temporal derivatives, resulting in 72 and 144 dimensional feature vectors, respectively. The FF features were obtained by filtering the logFBEs as described in Section 4. The logFBE features were obtained by passing the short-time magnitude spectra, obtained by applying the FFT, through the mel-spaced filter-bank analysis with 26 channels. The frame length and shift were 32 ms and 10 ms, respectively, and both pre-emphasis and Hamming window were applied to each frame. Masks employed in the MFT model were estimated as described in Section 2.2 for the static features; masks for the derivative features were set equal to static-feature masks.

For both feature representations, a continuous-observation left-to-right HMM with 3 states (no skip allowed) was used to model each phoneme; the pdf at each state was modeled with twenty-four Gaussian mixture components with diagonal covariance matrices. The HMMs were trained based on the provided scripts [16] using HTK [17]. An in-house Viterbi decoder developed in MATLAB was used for all recognition tests.

5.3. Experimental results using the oracle mask

Recognition accuracy results obtained by the FF and combined FF+logFBE features on the clean testset 1 were 87.5% and 89.8%, respectively.

Recognition results obtained by employing the FF features for noisy testsets when using the oracle mask are presented in Table 1, where ‘Marg’ and ‘Marg MP’ denote the marginalisation MFT without and with the mask-probability incorporated, respectively. It can be seen that the incorporation of the mask-probability significantly improved the recognition accuracy.

Table 1: Recognition accuracy [%] results obtained by the MFT-based ASR without and with the mask-probability incorporated when employing the frequency-filtered features (FF).

<table>
<thead>
<tr>
<th>Method</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS2</td>
<td>49.2</td>
</tr>
<tr>
<td>TS3</td>
<td>25.3</td>
</tr>
<tr>
<td>TS4</td>
<td>22.7</td>
</tr>
<tr>
<td>TS5</td>
<td>39.3</td>
</tr>
<tr>
<td>TS6</td>
<td>33.9</td>
</tr>
<tr>
<td>TS7</td>
<td>49.7</td>
</tr>
</tbody>
</table>

Table 2 provides recognition results obtained by using the combined FF+logFBE representation. Notation is the same as described above with the ‘MargBound’ denoting the bounded marginalisation MFT. Comparing the ‘Marg’ results from Table 1 and Table 2, it can be seen that the combined FF+logFBE features achieved much better results than those of the FF features alone in all cases. Comparing the ‘Marg’ and ‘MargBound’ results in Table 2, the bounded marginalisation, which has not been directly possible to use with the FF features, provided further improvement for each noise set. The incorporation of the mask-probability in both marginalisation and bounded-marginalisation resulted in improved performance. The improvement is especially high for testset 4 and 5.

The results presented above demonstrate that having an accurately estimated mask (i.e., a mask close to the oracle mask), the incorporation of the mask-probability could provide considerable performance improvement.
Table 2: Recognition accuracy [%] results obtained by the MFT-based ASR without and with the mask-probability incorporated when employing the combined FF+logFBE features.

<table>
<thead>
<tr>
<th>Method</th>
<th>TS2</th>
<th>TS3</th>
<th>TS4</th>
<th>TS5</th>
<th>TS6</th>
<th>TS7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marg</td>
<td>53.4</td>
<td>47.4</td>
<td>32.0</td>
<td>30.7</td>
<td>46.1</td>
<td>50.5</td>
</tr>
<tr>
<td>Marg MP</td>
<td>54.9</td>
<td>51.0</td>
<td>39.3</td>
<td>39.6</td>
<td>46.9</td>
<td>55.7</td>
</tr>
<tr>
<td>MargBound</td>
<td>54.9</td>
<td>50.5</td>
<td>35.7</td>
<td>33.6</td>
<td>48.7</td>
<td>53.9</td>
</tr>
<tr>
<td>MargBound MP</td>
<td>57.3</td>
<td>54.4</td>
<td>41.9</td>
<td>41.2</td>
<td>49.5</td>
<td>59.9</td>
</tr>
</tbody>
</table>

5.4. Experimental results using the onset mask

This section presents results obtained by using an automatic mask estimation based on a simple onset detection method. Results obtained by employing the FF features and the combined FF+logFBE features without and with the incorporated mask-probability are presented in Table 3. Due to the noise-stationarity assumptions inherent in the employed mask estimation method, results are presented for testset 4 only. The achieved results demonstrate that the use of the combined FF+logFBE features and incorporation of the mask-probability provided improvements consistent with the use of the oracle mask.

Table 3: Recognition accuracy [%] results for testset 4 obtained by the MFT-based ASR without and with the mask-probability incorporated when employing the estimated onset mask.

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature representation</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FF</td>
<td>TS2</td>
</tr>
<tr>
<td>Marg</td>
<td>16.7</td>
<td>18.2</td>
</tr>
<tr>
<td>Marg MP</td>
<td>21.9</td>
<td>25.5</td>
</tr>
<tr>
<td>MargBound</td>
<td>-</td>
<td>26.3</td>
</tr>
<tr>
<td>MargBound MP</td>
<td>-</td>
<td>28.7</td>
</tr>
</tbody>
</table>

6. Conclusion

In this paper, we presented a novel method that models the mask within the missing-feature ASR system. The mask model was estimated by a separate training procedure for each mixture at each HMM state. We also explored on feature representation within the missing-feature ASR system. Experimental evaluations were performed on the Consonant Challenge corpus. The effectiveness of the proposed methods was first demonstrated using the oracle masks for all testset noisy conditions. Experimental results demonstrated considerable recognition accuracy improvements by employing the combined FF+logFBE features and incorporation of the mask-probability. Last evaluations were performed on testset 4 by using an automatic mask estimation based on a simple onset detection method and achieved results showed that the methods investigated provided performance improvements consistent with the use of the oracle mask.

7. Acknowledgement

This work was supported by UK EPSRC grant EP/D033659/1.

8. References