N-best Based Stochastic Mapping on Stereo HMM for Noise Robust Speech Recognition

Xiaodong Cui\textsuperscript{1}, Mohamed Afify\textsuperscript{2} and Yuqing Gao\textsuperscript{1}

IBM T. J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY, 10598, USA\textsuperscript{1}
Orange Lab, Smart Village, Cairo, Egypt\textsuperscript{2}

Emails: cuix@us.ibm.com, mohamed_afify2001@yahoo.com, yuqing@us.ibm.com

Abstract

In this paper we present an extension of our previously proposed feature space stereo-based stochastic mapping (SSM). As distinct from an auxiliary stereo Gaussian mixture model in the front-end in our previous work, a stereo HMM model in the back-end is used. The basic idea, as in feature space SSM, is to form a joint space of the clean and noisy features, but to train a Gaussian mixture HMM in the new space. The MMSE estimation, which is the conditional expectation of the clean speech given the sequence of noisy observations, leads to clean speech predictors at the granularity of the Gaussian distributions in the HMM model. Because the Gaussians are not known during decoding, N-best hypotheses are employed. This results in a clean speech predictor which is a weighted (by posteriors) sum of the estimates from different Gaussian distributions. In experimental evaluation of the proposed method on the Aurora 2 database it gives better performance over the MST model, particularly, about 10%-20% relative improvement under unseen noise conditions.

Index Terms: speech recognition, noise robust, stochastic mapping, stereo HMM, word graph

1. Introduction

Noise robustness becomes crucial when a speech recognition system is deployed in real-world applications. In recent years, the IBM multilingual real-time automatic speech-to-speech translator (MASTOR) \cite{1}\cite{2} has been targeting its deployment in military conditions as part of the DARPA Transtac project. The automatic speech recognition component in the speech-to-speech translation system is demanded to be robust to the environment, especially with military noise, to function properly in real world conditions. Therefore, accomplishing noise robust speech recognition is a fundamental and important issue to the success of the translation system in this scenario.

In recent work we proposed a feature compensation algorithm that uses stereo data and is, thus, referred to as stereo-based stochastic mapping (SSM) \cite{3}. When integrated with the speech recognition front-end of the IBM speech-to-speech translation system it showed significant performance gains, especially in noisy real-field evaluations. The basic idea of SSM is to concatenate the clean and noisy channels of the stereo data, in any desired way, to form a new augmented joint space. A Gaussian mixture model (GMM) is then built in this new space to represent the joint distribution of the clean and noisy data. During decoding, the observed noisy features together with the joint GMM model are used to estimate the clean features. Two formulations of the estimation were presented. One is iterative and uses the maximum a posteriori (MAP) criterion \cite{3} while the other relies on minimum mean square error estimation (MMSE) \cite{4}.

Both formulations led to significant performance gains in noisy conditions and were shown to be mathematically related to other piece-wise linear transformations, such as SPLICE \cite{5}, stochastic vector mapping (SVM) \cite{6} and FMLLR \cite{8}.

In this paper we consider a natural extension of the above idea by linking the joint (clean-noisy) distribution to the recognition model instead of using a separate GMM. In other words, after the concatenation of the clean and noisy features, a model that has the same structure as the recognition Gaussian mixture HMM is constructed. The model estimated in the joint space is referred to as stereo HMM in this paper. For each Gaussian, this model will have both the clean and noisy distributions as its marginals and will have a correlation component that allows the estimation of the clean features from the noisy features during decoding. Linking the correlation to the recognition model is expected to offer better resolution, in the clean feature estimation, than a general GMM. Another advantage of building a stereo HMM is that the whole sequence of features is taken into account for clean feature estimation. This is in contrast to the frame posterior weighting used in the GMM case. This point will be discussed in more detail in the paper. One problem with the above formulation is that the recognition hypotheses are needed to perform the clean feature estimation, and we employ the N-best candidates to overcome this limitation.

The remainder of the paper is organized as follows. In Section 2, we give the mathematical formulation of N-
best based SSM and show how feature compensation and decoding are performed in the stereo HMM framework. The discussion is limited to MMSE estimation due to its non-iterative nature but extension to MAP based SSM is possible at the expense of additional complexity. Experimental results are presented in Section 3 and summary and conclusions are provided in Section 4.

2. Mathematical Formulation

This section first formulates a joint model of the clean and noisy features then shows how this model is used to perform clean features estimation in a stereo HMM framework.

Denote a set of stereo features as \( \{(x_m, y_m)\} \), where \( x \) is the clean speech feature vector, \( y \) the corresponding noisy speech feature vector, and \( m \) an index that denotes the vector number. In the most general case, \( y_m \) is \( L_n \) concatenated noisy vectors, and \( x_m \) is \( L_c \) concatenated clean vectors. Define \( z \equiv (x, y) \) as the concatenation of the two channels, where the subscript \( m \) has been removed for convenience. The concatenated feature vector \( z \) can be viewed as a new feature space where a Gaussian mixture HMM model can be built. In the general case, when the feature space has dimension \( M \), the new concatenated space will have a dimension \( M (L_c + L_n) \). An interesting special case that greatly simplifies the problem arises when only one clean and noisy vectors are considered, and only the correlation between the same clean and noisy feature vectors are taken into account. This reduces the problem to a space of dimension \( 2 M \) with the covariance matrix of each Gaussian having the diagonal elements and the entries corresponding to the correlation between the same clean and noisy feature element, while all other covariance values are zeros.

Training of the above Gaussian mixture HMM will lead to the transition probabilities between states, the mixture weights, and the means and covariances of each Gaussian. The mean and covariance of the \( k^{th} \) component of state \( i \) can be partitioned as

\[
\begin{align*}
\mu_{z,i,k} &= \left( \begin{array}{c}
\mu_{x,i,k} \\
\mu_{y,i,k}
\end{array} \right) \\
\Sigma_{zz,i,k} &= \left( \begin{array}{cc}
\Sigma_{xx,i,k} & \Sigma_{xy,i,k} \\
\Sigma_{yx,i,k} & \Sigma_{yy,i,k}
\end{array} \right)
\end{align*}
\]

where subscripts \( x \) and \( y \) indicate the clean and noisy speech features respectively.

For the \( k^{th} \) component of state \( i \), given the observed noisy speech feature \( y \), the MMSE estimate of the clean speech \( x \) is given by \( E[x|y, i, k] \). Since \( (x, y) \) are jointly Gaussian, the expectation is known to be

\[
E[x|y, i, k] = \mu_{x,y,i,k} - \sum_{y,i,k} \Sigma^{-1}_{yy,i,k}(y - \mu_{y,i,k})
\]

where \( \mu_{x,y,i,k} \) is a scaling factor and \( \Sigma^{-1}_{yy,i,k} \) is the posterior probability of staying at mixture component \( k \) of state \( i \) given the feature sequence \( y_1^T \) and hypothesis \( H \). This posterior can be calculated by the forward-backward algorithm on the hypothesis \( H \). The expectation term is calculated using Eq. 3. \( p(H|y_1^T) \) is the posterior probability of the hypothesis \( H \) and can be calculated from the N-best list as follows:

\[
p(H|y_1^T) = \frac{p(y_1^T|H)^{\nu} p(H)^{\nu}}{\sum_{j} p(y_1^T|H_j)^{\nu} p(H_j)^{\nu}}
\]

where the summation in the denominator is over all the hypotheses in the N-best list, and \( \nu \) is a scaling factor that need to be experimentally tuned.

By comparing the estimation using the stereo HMM in Eq. 5 with that using a GMM in the joint feature space [4] as shown in Eq. 7,
\[ \hat{x}_t = \sum_k p(k|y_t) E[x_t|k, y_t] \quad (7) \]

we can find out the difference between the two estimates. In Eq. 5, the estimation is carried out by weighting the MMSE estimate at different levels of granularity including Gaussians, states and hypotheses. Additionally, the whole sequence of feature vectors, \( y^T_t = (y_1, y_2, \ldots, y_T) \), has been exploited to denoise each individual feature vector \( x_t \). Therefore, a better estimation of \( x_t \) is expected in Eq. 5 over Eq. 7.

Fig. 1 illustrates the whole process of the proposed noise robust speech recognition scheme on stereo HMM. First of all, a traditional HMM is built in the joint (clean-noisy) feature space, which can be readily decomposed into a clean HMM and a noisy HMM as its marginals. For the input noisy speech signal, it is first decoded by the noisy marginal HMM to generate a word graph and also the N-best candidates. Afterwards, the MMSE estimate of the clean speech is calculated based on the generated N-best hypotheses as the conditional expectation of each frame given the whole noisy feature sequence. This estimate is a weighted average of Gaussian level MMSE predictors. Finally, the obtained clean speech estimate is re-decoded by the clean marginal HMM in a reduced searching space on the previously generated word graph.

A word graph based feature enhancement approach was investigated in [7] which is similar to the proposed work in the sense of two pass decoding using word graph. In [7], the word graph is generated by the clean acoustic model on enhanced noisy features using signal processing techniques and the clean speech is actually “synthesized” from the HMM Gaussian parameters using posteriori probabilities. In this paper the clean speech is estimated from the noisy speech based on the joint Gaussian distributions between clean and noisy features.

3. Experimental Results

The proposed method is evaluated on the Sets A and B of the Aurora 2 database. There are four types of noise in the training set which include subway, babble, car and exhibition noise. The test set A has the same four types of noise as the training set while set B has four different types of noise, namely, restaurant, street, airport and station. For each type of noise, training data are recorded under five SNR conditions: clean, 20 dB, 15 dB, 10 dB and 5 dB while test data consist of six SNR conditions: clean, 20 dB, 15 dB, 10 dB, 5 dB and 0 dB. There are 8440 utterances in total in the training set with the four types of noise contributed by 55 male and 55 female speakers. For the test set, each SNR condition of each noise type consists of 1001 utterances leading to 24024 utterances in total from 52 male and 52 female speakers.

Word based HMMs are used, with each model having 16 states and 10 Gaussian distributions per state. The original feature space is of dimension 39 and consists of 12 MFCC coefficients, energy, and their first and second derivatives. In the training set, clean features and their corresponding noisy features are spliced together to form the stereo features. Thus, the joint space has dimension 78. First, a clean acoustic model is trained on clean features on top of which single-pass re-training is performed to obtain the stereo acoustic model where the correlation between the corresponding clean and noisy components is taken into account. A separate multi-style trained (MST) model is also constructed in the original space to be used as a baseline. The results are shown in Tables 1-3. Both the MST model and the stereo model are trained on the mix of four types of training noise.

A word graph, or lattice, is constructed for each utterance using the noisy marginal of the stereo HMM and converted into an N-best list. Different sizes of the list were tested and results for lists of sizes 5, 10 and 15 are shown in the tables. Hence, the summation in the denominator of Eq. 6 is performed over the list, and different values (1.0, 0.6 and 0.3) of the weighting \( \nu \) are evaluated (denoted in the parentheses in the tables). The language model probability \( p(H) \) is taken to be uniform for this particular task. The clean speech feature is estimated using Eq. 5. After the clean feature estimation, it is rescored using the clean marginal of the stereo HMM on the word graph. The accuracies are presented as the average across the four types of noise in each individual test set.

From the tables we observe that the proposed N-best based SSM on stereo HMM performs better than the MST model especially for unseen noise in Set B and at low SNRs. There is about 10%-20% word error rate (WER) reduction in Set B compared to the baseline MST model. It can be also seen that there is little influence for the weighting factor. This might be due to the uniform language model used in this task but might change for other scenarios. By increasing the number of N-best candidates...
in the estimation, the performance increases but not significantly.

4. Summary and Conclusions

In this paper we presented an extension of our previously proposed stereo-based stochastic mapping (SSM) in front-end feature space to back-end HMM model space. Instead of an auxiliary separate Gaussian mixture model in the previous work, an HMM model is built on stereo features. The basic idea, as in SSM, is to form a joint space of the clean and noisy features, but to train a Gaussian mixture HMM in the new space. For the MMSE formulation this leads to clean speech predictors at the granularity of the Gaussian distributions. These predictors are the conditional expectations of the clean speech given the sequence of noisy observations. Because the Gaussians are not known during decoding we use N-best hypotheses. As a result, the final clean speech predictors are a weighted sum of the estimates from different Gaussian distributions.

In the recognition stage, a word graph and N-best hypotheses are first generated by the noisy marginal HMM of the stereo HMM and the MMSE estimation of the clean speech is performed based on the N-best hypotheses. Finally the estimated clean speech is decoded by the clean marginal HMM on the word graph.

In the experimental evaluation of the proposed method on the Aurora 2 database it yields better performance over the MST model, especially in the unseen noise and low SNR cases where 10%-20% WER reduction is observed.

5. References


