Combining Noise Compensation and Missing-Feature Decoding for Large Vocabulary Speech Recognition in Noise

Jianhua Lu, Ji Ming, Roger Woods

The Institute of Electronics, Communications and Information Technology
Queen’s University Belfast
{jlv01,j.ming,r.woods}@qub.ac.uk

Abstract

In this paper we propose a combination of noise compensation and missing-feature decoding for large-vocabulary speech recognition in noisy environments. Two approaches for noise compensation have been studied. These are noise training and vector Taylor series expansion, aiming to compensate white Gaussian noise at various levels. This is followed by subband missing-feature decoding to reduce the model/data mismatch. The proposed approach requires little knowledge about the noisy environments. The Aurora 4 corpus is used for the experiments. Better results are obtained by the new approach over a multi-condition baseline system.

Index Terms: noise compensation, missing-feature decoding, noise robustness, speech recognition, Aurora 4

1. Introduction

Clean speech recognition has been proved to have very high accuracy. For isolated word recognition, the word error rate (WER) achieved can be as low as 2% or less. For large vocabulary continuous speech recognition (LVCSR), the WER can be reduced to within 10%. However, when such systems are used in acoustic environments with background noise, the recognition performances degrade dramatically as the signal-to-noise ratio (SNR) decreases [1]–[4]. Even models trained from a wide range of noisy environments cannot rule out the existence of mismatches when applied to an unseen environment. There have been a number of techniques for dealing with mismatched training and testing, based on the learning of the statistics of noise. The learnt statistics can be used to compensate the effects of noise thereby reducing the training and testing mismatch. Examples of these noise compensation techniques include parallel model combination (PMC) [1], which uses equation-based log-normal approximation or data-driven model retraining to reduce the mismatch; vector Taylor series (VTS) [4], [2], which represents a different type of approximation to the nonlinear mismatch function; and SPLICE or uncertainty decoding [5], [6], which uses stereo or adaptive training data to learn the mismatch corrections. Additionally, missing-feature techniques have been used to reduce the training and testing mismatch by exploiting the time-frequency regions which provide reliable information for the recognition [7]–[9]. Previous experiments have indicated that missing-feature methods are capable of dealing with non-stationary noise since reliable features may be detected frame by frame [9], while the other noise compensation techniques usually need longer period of observations to learn the noise statistics.

In this paper, we investigate the combination of noise compensation and missing-feature techniques. We will show that the combination could lead to improved robustness when dealing with more sophisticated and non-stationary noise, in comparison to the noise compensation and missing-feature techniques operating in isolation. Specifically, we have extended the work in [9] in three aspects: (1) incorporating the VTS technique into the system as an alternative to white noise training for noise compensation; this makes it possible to remove the need for retraining the models for noise compensation and therefore existing recognition systems trained on clean data can be adapted efficiently to noisy speech recognition; (2) modifying the algorithm to detect the reliable subband features for improving the performance; and (3) applying the new system from previous connected word recognition to LVCSR, in particular, on the Aurora 4 database.

2. Proposed System

The proposed recognition system framework is shown in Fig.1, which consists of a noise compensated, subband missing-feature decoder for LVCSR in noise. The figure shows two paths to reach the decoder. On the left is the approach described in [9], which realizes the noise compensation by training the acoustic models using artificially corrupted training data using white noise at different SNRs. The white noise corrupted training aims to compensate white noise components in the testing data. The mismatch between the white noise and the actual testing noise will be reduced in the missing-feature decoding by focusing the recognition on the matched subbands. Therefore the system effectively requires no prior knowledge about the noise. On the right is an alternative approach to this goal studied in this paper, in which the noise compensation is realized by using the VTS technique, applied to subband acoustic models. Without assuming prior information about the testing noise, we assume white Gaussian noise and generate the Taylor series at different SNRs to compensate a range of possible noise. The mismatch between the assumed noise and the actual testing noise will be reduced in the missing-feature decoding by focusing on the matched subbands, as in the above approach. The new approach is of great interest not only because it has the potential to deal with a range of realistic noises without requiring prior information, but also because it can be applied to existing recognition systems if no training data available. Fig.1 also shows a possible conversion from a fullband acoustic model to the new subband system which will be discussed later.

3. Noise Compensation

The noise compensated acoustic model, represented by the probability distribution \( p(Y | s) \) of noisy speech frame \( Y \) con-
dictioned on state \( s \), can be expressed as
\[
p(Y|s) = \sum_{l=1}^{L} P(l|s)p(Y|l, s) \tag{1}
\]
where \( L \) is the number of mixtures for state \( s \), corresponding to the number of SNR resolutions in the compensation, and \( P(l|s) \) is the mixture weight, corresponding to the prior probability for SNR level \( l \) within state \( s \). In our system, each frame \( Y = (y_1, y_2, ..., y_K) \) is a set of \( K \) subband features with \( y_k \) being the feature for subband \( k \). Furthermore, assume that the subbands are independent with each other.

Model (1) can be obtained by adding white noise to the clean training frame \( X \) at \( L \) SNRs and then train the corresponding probability distribution \( p(Y|l, s) \) at each SNR level \( l \) (i.e., the left path). Alternatively, it can also be obtained from clean model \( p(X|s) \) by expanding the noisy frame \( Y \), which is a nonlinear function of the clean frame \( X \) and noise \( N \), into a Taylor series at \( L \) SNR levels and taking the first order approximation at each SNR level (i.e., the right path) [2]. The compensated mean vector and covariance matrix at SNR level \( l \) can be expressed as (we have omitted the state index for simplicity)
\[
\mu_{Y,l} = A_l \mu_X + b_l \tag{2}
\]
\[
\Sigma_{Y,l} = A_l \Sigma_X A_l^T
\]
where \( \mu_X \) and \( \Sigma_X \) are the mean vector and covariance matrix of the clean models, and \( A_l \) and \( b_l \) are the first-order Taylor coefficients for SNR level \( l \), which can be expressed as
\[
A_l = \frac{\partial g(X, N)}{\partial X} \bigg|_{X=X_0, N=N_l} \tag{3}
\]
\[
b_l = (I - A_l) X_0 + g(X_0, N_l) \tag{4}
\]
where \( g(\cdot) \) is the mismatch function for additive noise, \( X_0 \) is the expansion point, and \( N_l \) is the Gaussian compensation noise at SNR level \( l \) calculated against \( X_0 \). In our experiments, the global mean of the clean training data is used as \( X_0 \). Similar compensations are also applied to the delta and delta-delta.

4. Missing-Feature Decoding

In decoding, for each testing frame, only the subband features that are matched by the compensation model are included in the calculation of the likelihood of the frame. The mismatched subbands are deemphasized thereby reducing the model/data mismatch. The matched subbands at each SNR level can be selected by maximizing the joint posterior probability of the state and SNR level given the frame [9], i.e.,
\[
Y_{l,s} = \text{arg max}_{Z \in \mathbb{C}} P(l, s|Z) \tag{5}
\]
where \( Y_{l,s} \) represents the optimal set of subbands from \( Y \) that match the SNR condition \( l \) in state \( s \), in terms of maximizing the joint posterior probability. Following Bayes’ rules the posterior probability \( P(l, s|Z) \) can be expressed as
\[
P(l, s|Z) = \frac{p(Z|l,s)p(l,s)}{p(Z)} \tag{6}
\]
where \( p(l,s) \) is the prior probability of the state \( s \) and SNR level \( l \) and, in our previous work, \( p(Z) \) is approximated by
\[
p(Z) \approx \sum_{l,s} p(Z|l,s)p(l,s) \tag{7}
\]
This approximation is appropriate when the variation of the testing noise can be fully covered by the \( L \) levels of compensation noise. In this paper, we extend (7) by including a factor to account for the noise variations that are not covered by the \( L \) noisy training conditions. The new approximation is expressed below:
\[
p(Z) \approx \sum_{l,s} p(Z|l,s)p(l,s) + p(Z|\phi) \tag{8}
\]
In (8), the first term is the same as in (7), and the second term is an added model to account for the probability of subband feature \( Z \) given undiscovered model \( \phi \).

There can be a number of methods to approximate \( p(Z|\phi) \). The following is merely one of the choices, for \( Z = (z_1, z_2, ..., z_M) \) being a subset of \( M \) subbands from \( Y \):
\[
p(Z|\phi) \approx \prod_{m=1}^{M} \max_{l,s} p(Z|l,s)p(z_m|l,s) \tag{9}
\]
where \( p(z_m|l,s) \) is the likelihood of subband feature \( z_m \) in state \( s \) with SNR level \( l \). In (9), we break the correlation between the subbands and reconstruct a model by combining subbands from different states and SNR levels, as an approximation to \( p(Z|\phi) \).

5. Experimental Results

The Aurora 4 database is used for the experiments. The database is dedicated to evaluating and comparing the robustness of LVCSR algorithms. The database provides with two types of microphones, wv1 and wv2, and 6 types of real noises: car, babble, restaurant, street, airport and train, with a total of 14 test sets available for both 8 kHz and 16 kHz sampling frequencies. The noises are added to the 16 kHz sampled clean speech at different SNRs from 5 dB to 15 dB. Some noises are fairly stationary like the car noise. Others contain non-stationary components like the street and airport noises. In this
paper we only use the 16 kHz and wv1 test sets. For the training data, 7138 clean utterances are provided. We have first created a clean context independent recognition system with 28761 models which consists of 3238 states with 16 mixtures in each state. Next, we have implemented two methods for noise compensation: (1) the method based on training the acoustic models using noisy data with simulated noise at different SNRs (i.e., the left path in Fig.1), and (2) the method based on the VTS expansion for compensating white Gaussian noise at different SNRs (i.e., the right path in Fig.1). For the first method, Gaussian white noise is added at 20, 15, 10, 5, 0 dB, respectively. These noisy data, plus the original clean training data, make a total of 42828 training utterances which are used to train the compensated models. The added Gaussian noise is filtered using a low pass filter with a cutoff frequency of 2 kHz, to simulate the high-frequency rolloff effect of most realistic noises. For the second method, the VTS expansion (i.e., (2)–(4)) is calculated at each of the same SNR levels as in the first method, and the compensated models covering these SNR conditions are then created from the clean models by incorporating these expansions. Note that in both methods, we only use white Gaussian noise to compensate the models. That is, we assume no prior information about the actual corrupting noise. The mismatch in our compensations will be dealt with by the missing-feature decoding scheme, as described in Section 4.

For implementing the missing-feature decoding, we use subband features to represent each frame. The subband features are derived from the decorrelated log filter-band energies (FBE) [10] with 20 mel-scaled filters. The average over the full 20 filter energies is appended to the subband feature vector. Delta and delta-delta coefficients of the static features are also added to the frame vector and treated as additional subband features. We refer to [3] as a baseline system for comparison for both clean and multi-condition tests. Following recommendations from the Aurora 4 database, 166 out of 330 utterances of the ARPA test set are used in the experiments. The baseline system trained on clean data obtained a WER of 11.2% for clean testing data. We use the same beam pruning as in the baseline system.

Table 1 shows the WERs of three decoding algorithms based on clean acoustic models trained using the 7138 clean training utterances. In the table, “Product” stands for the system which uses the full set of subbands in the decoding and thus the likelihood of each frame equals the product of the likelihoods of each individual subbands; “MF” stands for the missing-feature decoding algorithm which selects an optimal set of subbands based on (5) and uses (6) to calculate the denominator \( p(Z) \); and “Extended MF” calculates the denominator using the extended equation (8).

As indicated in Table 1, both missing-feature algorithms have improved the recognition accuracy, reducing the WER from 53.2% for the product model to 38.8% and 37.5%, respectively, for the MF and extended MF models. On average, the extended MF algorithm performs slightly better than the MF algorithm, and particularly in the car and babble noise cases.

Table 2 shows the results based on the compensated acoustic models trained using simulated white noise. In comparison to the product results in Table 1, we have experienced a slight increase in WER for the clean testing data as expected because the noisy training increases the model variances. However, the compensation leads to significant WER reductions for the noisy testing conditions, reducing the average WER for all the three decoding algorithms. The importance of combining missing-feature decoding with noise compensation is made clear by the results. This combination improves the robustness against the training and testing mismatch. Note that white noise training plus missing-feature decoding results in a comparable WER (14.5%) of clean test with clean training/clean test (13.7%).

Table 3 quotes the average WER obtained by the baseline models [3] trained using the multi-condition training data, which contain noise samples from the testing conditions in the same channel, from the Aurora 4. Our models based on the combination of noise compensation and missing-feature decoding shown in Table 2 compares favorably with this multi-condition baseline, in terms of reducing the overall WER by 12.0%.

Table 4 shows a comparison between the two noise compensation approaches, one based on training in white noise (i.e., Table 2) and the other on extending the clean models using the VTS expansion to model white noise. Both approaches use missing-feature decoding to reduce the model/testing data mismatch. Two noisy cases are included in the table for illustration. The VTS compensation performs similarly to the noise training compensation for the clean testing condition. It outperforms the noise training compensation for the car noise condition, reducing the WER from 23.4% to 19.2% with the use of missing-feature decoding. However, it performs worse than the noise training compensation approach for the rest of the noisy testing conditions. This may be due to the inaccuracy of simulating re-

<table>
<thead>
<tr>
<th>Test condition</th>
<th>Decoding algorithm</th>
<th>Product</th>
<th>MF</th>
<th>Extended MF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td></td>
<td>16.5</td>
<td>14.0</td>
<td>13.7</td>
</tr>
<tr>
<td>Car</td>
<td></td>
<td>58.4</td>
<td>22.3</td>
<td>18.4</td>
</tr>
<tr>
<td>Babble</td>
<td></td>
<td>58.8</td>
<td>45.7</td>
<td>39.5</td>
</tr>
<tr>
<td>Restaurant</td>
<td></td>
<td>53.8</td>
<td>53.6</td>
<td>54.3</td>
</tr>
<tr>
<td>Street</td>
<td></td>
<td>62.5</td>
<td>50.8</td>
<td>51.5</td>
</tr>
<tr>
<td>Airport</td>
<td></td>
<td>56.9</td>
<td>37.0</td>
<td>35.5</td>
</tr>
<tr>
<td>Train</td>
<td></td>
<td>65.5</td>
<td>47.9</td>
<td>49.8</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>53.2</td>
<td>38.8</td>
<td>37.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test condition</th>
<th>Decoding algorithm</th>
<th>Product</th>
<th>MF</th>
<th>Extended MF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td></td>
<td>19.6</td>
<td>15.3</td>
<td>14.5</td>
</tr>
<tr>
<td>Car</td>
<td></td>
<td>35.9</td>
<td>23.4</td>
<td>19.1</td>
</tr>
<tr>
<td>Babble</td>
<td></td>
<td>33.6</td>
<td>22.6</td>
<td>23.2</td>
</tr>
<tr>
<td>Restaurant</td>
<td></td>
<td>36.8</td>
<td>27.1</td>
<td>27.1</td>
</tr>
<tr>
<td>Street</td>
<td></td>
<td>33.2</td>
<td>24.4</td>
<td>24.6</td>
</tr>
<tr>
<td>Airport</td>
<td></td>
<td>34.4</td>
<td>21.0</td>
<td>20.3</td>
</tr>
<tr>
<td>Train</td>
<td></td>
<td>35.4</td>
<td>26.5</td>
<td>24.8</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>32.7</td>
<td>22.9</td>
<td>21.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test condition</th>
<th>Clean</th>
<th>Noisy</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>18.1</td>
<td>26.0</td>
<td>24.9</td>
</tr>
</tbody>
</table>
alistic noise with the first-order expansion against white noise. This needs to be studied further. An advantage of the proposed VTS compensation for simulated noise is that it offers a way of implementing noise compensation in existing systems without requiring training data, as discussed below.

A comparison between the proposed new approach and the baseline in decoding time is shown in the following Table 5.

Table 5: Comparison of baseline Viterbi and missing-feature decoding in CPU time

<table>
<thead>
<tr>
<th>Method</th>
<th>Relative CPU time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Viterbi</td>
<td>1.0</td>
</tr>
<tr>
<td>MF decoding</td>
<td>1.0</td>
</tr>
</tbody>
</table>

6. Model Conversion

In this paper, subband features are used for each frame which are derived from the decorrelated log FBE. However, for most state-of-the-art speech recognition systems, a fullband MFCC front-end is used. A simple transformation may be applied to MFCC based models to obtain the corresponding subband models. Then, the VTS compensation can be applied to the obtained subband models thereby realizing noise compensation.

Assume a fullband MFCC consisting of \( M \) coefficients. These can be transformed to \( F \) log FBE using an inverse DCT. Uniformly dividing these \( F \) log FBE into \( K \) subbands, and applying a DCT to each subband, we can obtain a new set of subband MFCCs. Applying these transforms to the fullband models, we can obtain the corresponding subband models. Let \( C \in \mathbb{R}^{M \times F} \) be the fullband DCT matrix, and \( C_k \in \mathbb{R}^{M_x \times F_k} \) be the DCT matrix for subband \( k \), the subband model parameters can be derived as follows

\[
\mu_{sb} = \hat{C} C^{-1} \mu_{fb} \\
\Sigma_{sb} = \hat{C} C^{-1} \Sigma_{fb} C^{-T} \hat{C}^T
\]

(10)

where \( \mu_{sb}^T = \{ \mu_1^T, \mu_2^T, ..., \mu_K^T \} \) and \( \Sigma_{sb} = \text{diag}\{\Sigma_1, \Sigma_2, ..., \Sigma_K\} \).

\[
\hat{C} = \text{diag}\{C_1, C_2, ..., C_K\}
\]

(11)

The model for subband \( k \) is defined by \((\mu_k, \Sigma_k)\), while \((\mu_{fb}, \Sigma_{fb})\) represents the original fullband model.

We have examined the above model transformation by applying it to a fullband system taken from [11], which consists of 8000 states with 4 mixtures in each state. We have transformed the system into a 36-dimensional, 6-subband system with two static, delta and delta-delta coefficients in each subband. The new subband system is examined on the Nov’92 evaluation test set, consisting of 8 speakers with about 40 utterances from each speaker. All subbands, i.e., the product model, are used for the recognition. The WER of the original fullband system is 7.10%, while that of the transformed subband system is 9.94%, with 2.84% absolute errors being introduced by breaking the correlation between the frequency bands.

7. Conclusions

Missing-feature techniques can be incorporated into clean models to provide considerable robustness to partial noise corruption. However, for noises spreading across the whole time/frequency range, missing-feature decoding offers less significant improvement. This paper demonstrated, via experiments on the Aurora 4 database, that combining noise compensations with missing-feature decoding could significantly reduce the error rate for noisy data while not seriously damaging the accuracy for clean test.

8. Acknowledgments

This work is supported by the UK EPSRC under Grant No. EP/D048605/1.

9. References


