Modeling the Mixtures of Known Noise and Unknown Unexpected Noise for Robust Speech Recognition

Ji Ming, Peter Jancovic, Philip Hanna, Darryl Stewart

School of Computer Science
Queen's University of Belfast, Belfast BT7 1NN, UK
j.ming, p.jancovic, p.hanna, dw.stewart@qub.ac.uk

Abstract

Real-world noise may be a mixture of known or trainable noise and unknown unexpected noise. This paper investigates the combination of the conventional noise-reduction techniques with the probabilistic union model to deal with this type of mixed noise for robust speech recognition. In particular, we have developed a multi-environment system to remove the known or trainable acoustic mismatch across different environments. The novelty of this system, in contrast to other multi-environment models, is that the acoustic model for each environment is built upon the probabilistic union model, so that this system is also capable of accommodating further unknown unexpected noise within a specific environment. We have tested the new system for connected digit recognition in different environments, each involving an environment-specific noise and some unknown untrained noise. The results indicate that the new system offers significantly improved performance for the environments involving unknown additional noise, in comparison to a baseline multi-environment system.

1. Introduction

Speech recognition performance is known to degrade dramatically when a mismatch occurs between training and testing environments. This mismatch can be due to a number of factors, with ambient or communication-channel noise being one of the most prominent. Real-world noise may be modelled by a mixture of stationary noise and nonstationary or unknown noise. For stationary noise, with reasonably sufficient observations, it is possible to obtain an estimate of the noise characteristics. This knowledge may be used to reduce the mismatch between the model and test data caused by the stationary noise or noise of a known characteristic. Conventional techniques for this range from noise filtering to feature or model compensation (e.g. [1]). In particular, the model compensation techniques based on parallel model combination [2] or on multi-environment models [3] have been shown to be an effective approach to handle a wide range of noise variations. In addition to the stationary noise component, a real-world environment may also include some nonstationary or unknown unexpected noise components, which may be the leftover of an inaccurate model compensation, or an additional corruption occurring to the utterance. This type of noise may be handled by using robust acoustic features, or by using the missing feature method (e.g. [4]-[8]). Assuming that the noise affects only certain parts in the temporal-spectral feature space, the missing feature method suggests that the recognition may be based on information from the clean parts, by throwing away the noisy parts, or by making the noisy parts play a less significant role in recognition. This recognition is made possible due to redundancy in the temporal-spectral characteristics of speech. This method is of interest because there can be situations where removing the noise from the observation may prove difficult, due to the lack of sufficient knowledge about the noise. A better system may be a combination of these two methods, i.e., using the noise reduction technique to remove the noise with a known or trainable characteristic, and exploiting the redundancy in the speech signal to get round the noise with an unknown or time-varying nature. In this paper we investigate the combination of the model compensation technique with the probabilistic union model for this purpose. In particular, we employ the multi-environment model technique to reduce the trainable acoustic mismatch across a variety of environments, and employ the probabilistic union model to accommodate further unknown unexpected corruption within each environment.

Because the missing feature method is not aimed to clear the noise from the noisy observation, it does not require a detailed knowledge of the noise. It only requires labeling every temporal-spectral feature as reliable or corrupt, for removing the unreliable features from recognition. Unfortunately, locating the corrupted features itself can be a difficult task, if there is no prior information about the noise. Mistakes in labeling the features can cause either a loss of reliable information, or an inclusion of unreliable information in the recognition process. As suggested in [6][7], the unreliable features may be identified by explicitly measuring the local signal-to-noise ratio (SNR), based on a running estimate of the local noise spectrum via spectral subtraction. This method performs well when the corrupting noise is stationary. But it may fail to produce accurate estimates in nonstationary noise or unknown noise [8], as in these conditions the assumption required for spectral subtraction is invalidated. To overcome this problem, we have recently proposed the probabilistic union model [9][10]. Unlike the missing feature method, the new model does not require the identity of the noisy data, instead, it combines the local temporal-spectral information based on the probability theory for the union of random events, to reduce the dependence of the model on information about the noise. This model improves upon the missing feature method in that it offers robustness against partial feature corruption, while requiring no information about the noise. In the following, we first describe the proposed multi-environment system, and then the probabilistic union model used for each environment. An evaluation of the system for noisy connected digit recognition is presented in Section 3.

2. The Multi-Environment Model

The proposed multi-environment system is shown in Fig. 1. As shown, the system has different acoustic models to match a variety of environments. For each specific environment, the corresponding acoustic model may be derived from a general-purpose (i.e. environment-independent) model by transforming the model parameters to better characterize the acoustic condition of the given environment. This transformation may be performed based on pre-trained parameter translation vectors associated with the environment, or based on parallel model combination assuming the availability of the underlying noise model. For some applications, the environment-specific model may also be obtained by directly training the model in the appropriate environment.

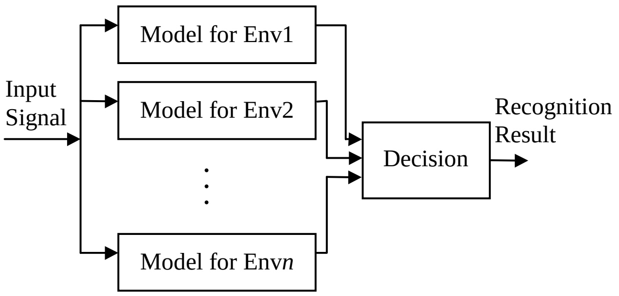

Figure 1: Schematic diagram of the multi-environment model

In recognition, given a segment of speech, there may be two ways to decide an appropriate model to perform recognition. Firstly, if the environment can be identified before recognition, then only the model matching the identified environment is used for recognition. An example of this method has been described in [3]. Alternatively, if there is a difficulty in identifying the environment before recognition, then all the models that have a potential to match the testing environment should be employed to perform recognition, with each model producing a recognition result. Based on these results, a final decision, based on some optimality criterion, is made to select the final recognition result.

The above multi-environment model is suited for dealing with acoustic mismatch due to changes in environment. Within each environment, the model effectively assumes either that the noise condition will remain the same as in the training stage, or that a full knowledge of the underlying noise can be available for accurate model compensation. These assumptions are generally not true for real-world situations. To enhance the capability of the model for dealing with unknown unexpected noise within each environment, we build the acoustic model for each environment based on the probabilistic union model, described below.

3. The Probabilistic Union Model

Assume that a test utterance is represented by a set of N temporal-spectral feature vectors $X = (x_1, x_2, ..., x_N)$, and P_i, $i = 1, 2, ..., N$, is the environment-matched probability distribution for each x_i. Further, assume that in addition to the known environment-specific noise, X may also be subjected to some *unknown* noise corruption, which cause some of the feature vectors x_i to deviate from the pre-trained environment model P_i. The idea of the missing feature method is that the acoustic mismatch due to the unknown noise can be effectively reduced by simply ignoring strongly affected feature vectors. However, because of the uncertainty of the noise, the identities of the corrupted feature vectors are unknown. The probabilistic union model is a method that can be used to select usable features from a given feature set, without requiring the identity of the corrupted features [9][10]. This capability makes it a suitable model for being used in each environment, to provide robustness to any further unknown unexpected noise corruption.

The probabilistic union model deals with the uncertainty of the corrupted feature vectors by combining the feature vectors x_i using the inclusive "or" (i.e. disjunction) operator. Let $P(X)$ be the probability of the whole feature set X. With the union model, this probability can be expressed in a general form

$$P(X) = P(\bigvee_{n_1 n_2 \cdots n_{N-M}} x_{n_1} x_{n_2} \cdots x_{n_{N-M}}) \qquad (1)$$

where the symbol $\vee$ represents the inclusive "or" combination, applied to combine all the $x_{n_1} x_{n_2} \cdots x_{n_{N-M}}$'s, which are each a subset of $(N - M)$ feature vectors within $(x_1, x_2, ..., x_N)$; M is called order of the union model, with a value within the range $0 \le M < N - 1$. For example, in the case with four feature vectors x_1, x_2, x_3, x_4, the union model probability $P(X)$ can take four possible forms, corresponding to order $M = 0, 1, 2$ and 3, respectively:

$$M\text{=0:} \ P(X) = P(x_1 x_2 x_3 x_4) \qquad (2)$$

$$M\text{=1:} \ P(X) = P(x_1 x_2 x_3 \vee x_1 x_2 x_4 \vee x_1 x_3 x_4 \vee x_2 x_3 x_4) \qquad (3)$$

$$M\text{=2:} \ P(X) = P(x_1 x_2 \vee x_1 x_3 \vee x_1 x_4 \vee x_2 x_3 \vee x_2 x_4 \vee x_3 x_4) \quad (4)$$

$$M\text{=3:} \ P(X) = P(x_1 \vee x_2 \vee x_3 \vee x_4) \qquad (5)$$

The above indicates that the traditional model, (2), which combines all the feature vectors using the "and" (i.e. conjunction) operator, is a special case of the union model with order $M = 0$. A union model of order M is suited for accommodating a maximum of M corrupted feature vectors without requiring their identity. To illustrate this, use the above example with order 2, assuming two corrupted feature vectors with unknown identity. The union probability $P(X)$ for order $M = 2$ can be approximated as

$$P(X) \approx P(x_1 x_2) + P(x_1 x_3) + P(x_1 x_4) + P(x_2 x_3) + P(x_2 x_4) + P(x_3 x_4) \qquad (6)$$

where we have omitted the terms corresponding to the joint probabilities between the $x_i x_j$'s for simplicity. As indicated in (6), the union model includes the probabilities of all possible combinations between two feature vectors, and thus it will include the probability for the remaining two "clean" feature vectors, providing correct information about the probability of X . The probability containing only the clean feature vectors should usually dominate the probability $P(X)$ over the correct model, because of small mismatch between the model and data. As such, recognition can be based on the union probability $P(X)$, and hence no information is required for the identity of the two corrupted feature vectors.

The above union model has been previously implemented within an HMM framework and applied to the combination of sub-band features [9], time-segment features and different types of feature streams [10], showing strong robustness to partial, unknown, time-varying noise corruption. In this paper we combine this model with the model compensation technique described above, to deal with the mixture of stationary or trainable noise and unknown unexpected noise.

4. Experiments

A multi-environment system as shown in Fig. 1 has been implemented; for each environment, an acoustic model based on the probabilistic union model was built to perform recognition. Specifically, we considered four different environments, i.e. clean, car, train and restaurant. For the environments involving noise (i.e. car, train and restaurant), the signal-to-noise ratio (SNR) is about 10 dB. For each environment, we assumed that we had a set of acoustic models trained in the environment, to match the known environmental characteristics. Alternatively, the knowledge about an environment may also be given as a model of the noise or as an extra set of training data in the new environment, so that environment-specific model may be derived by noise filtering, parallel model combination or parameter transformation.

The TIDigits connected digits database was used in the experiments. This database contains a total of 6196 test utterances for speaker-independent connected digit recognition. Each test utterance contained a string of 2, 3, 4, 5 or 7 digits, respectively, assuming no advance knowledge of the number of digits in an utterance. The speech was sampled at 8 kHz, and divided into frames of 256 samples. For each frame, we calculated five sub-band feature vectors and their corresponding delta vectors, resulting in a total of ten different feature streams for recognition. The sub-band feature vectors were calculated as follows. Firstly, a mel-scaled filter bank was used to estimate the log-amplitude spectra of each frame, then these filter-bank channels were grouped uniformly into sub-bands, and for each sub-band, the MFCCs were calculated to form the sub-band frame vector. In our experiments, each sub-band frame vector contained three MFCCs; so the total number of parameters for each frame is 10 streams × 3 coefficients per stream, or 30. The union model was trained to combine these ten feature streams, each stream corresponding to a feature vector x_n as shown in (1) (for a full description of the implementation of the union model, see [9][10]). For comparison, we also implemented a multi-environment system based on a set of baseline models, with each model trained for a specific environment as the union model and using the full-band feature vectors. Both the union model and baseline model were built on a continuous-observation HMM framework with eight mixtures for each state and ten states for each digit. In the multi-environment systems, no explicit environment identification was performed before recognition, so all the models for different environments were run in parallel and a final decision about the result was made at the end of the recognition process, based on a maximum probability criterion.

Firstly, we tested the union model and baseline model for all possible combinations between the training and testing environments, assuming no additional noise in the testing environment. Table 1 presents the results, showing the digit *string* accuracy obtained by the two models, respectively, for each training-testing environment combination. As shown in Table 1, both models achieved the best accuracy for matched environment training and testing. In particular, the baseline model achieved a string accuracy of 97.53% for the clean condition, which is comparable to that reported in [11], 97.06%. Importantly, Table 1 indicates that the union model is more robust to environment mismatch than the baseline model, due to the model's capability of handling additional corruption.

Table 1: Digit string accuracy (%) for the baseline model (top) and union model (bottom) for different training and testing conditions (i.e. clean, car, train and restaurant)

Testing Training	Clean	Car	Train	Rest.
Clean	97.53 97.19	45.77 49.55	81.21 83.38	51.90 57.26
Car	43.38 51.89	91.79 88.57	79.12 82.78	62.31 75.32
Train	88.99 90.62	75.53 79.74	94.75 94.24	65.62 79.95
Rest.	88.56 89.86	82.71 84.55	86.86 91.25	89.62 88.91

Table 2 presents the recognition results produced by the multi-environment systems based on the baseline model and union model, respectively. Within each system, no knowledge about the testing environment was assumed; all acoustic models were run in parallel and the result was selected based on the maximum probability. As can be seen, the two systems achieved a similar performance in all the testing conditions.

Table 2: Digit string accuracy (%) for the multi-environment systems based on the baseline model and union model, respectively, for different testing environments without additional noise

Testing condition	Clean	Car	Train	Rest.
Baseline	97.16	87.49	92.83	82.63
Union	96.19	86.93	92.90	82.94

We then tested the two multi-environment systems by assuming that in each testing environment, the speech utterances were also corrupted by some unknown additive noise which were not seen in the training stage. Three types of noise were considered as the additional noise: 1) a whistle, 2) a telephone ring, and 3) a bell. These noises were added, respectively, to the test utterances from each testing environment to simulate a further unknown unexpected corruption occurring to the utterances in that environment. The SNR of the additional noise was 10 dB, so the overall SNR, taking into account both the environment noise and the additional noise, was about 7 dB. Table 3 presents the recognition results, averaged over the three types of additional noise as described above. As shown in Tables 3, the multi-environment system based on the union model offers considerable performance improvements over the system based on the baseline model. As a summary, Fig. 2 shows a comparison between the two systems over all the test conditions described in this paper. As shown in Fig. 2, the two systems offer similar performance for trained environments without additional noise, but the union system offers significantly improved performance for the environments involving unknown additional noise.

Table 3: Digit string accuracy (%) for the multi-environment systems for different testing environments with unknown additional noise; the new testing environment with additional noise is denoted by "environment+"

Testing condition	Clean+	Car+	Train+	Rest.+
Baseline	58.65	51.18	33.62	34.01
Union	87.97	72.87	78.26	69.19

5. Summary

This paper described a combination of the traditional noise compensation technique with the probabilistic union model for dealing with a mixture of known noise and unknown unexpected noise. We have built a multi-environment system to reduce the trainable acoustic mismatch across a variety of environments; the system employs the probabilistic union model to accommodate further unknown noise within each specific environment. We have tested the new system for connected digit recognition for different environments, each environment also involving some unknown untrained noise. The experimental results show that the new system offered significantly improved performance for the mismatched environments, in comparison to a baseline multi-environment system.

This work is supported by UK EPSRC grant GR/M93734.

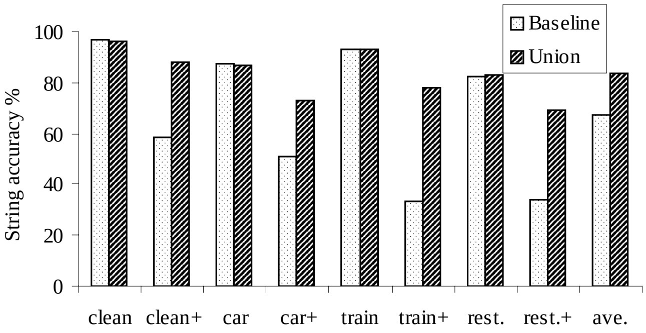

Figure 2: Summary of performance for environments without unknown additional noise (e.g. car) and with unknown additional noise (e.g. car+).

6. References

[1] Bellegarda, J. R., "Statistical techniques for robust ASR: review and perspectives", *Eurospeech'97*, pp. 33-36, 1997.

[2] Gales, M. J. F. and Young, S. J., "An improved approach to the hidden Markov model decomposition of speech and noise", *ICASSP'92*, pp. 233-236, 1992.

[3] Rahim, M., "A parallel environment model (PEM) for speech recognition and adaptation", *Eurospeech'97*, pp. 1087-1090, 1997.

[4] Lippmann, R. P. and Carlson, B. A., "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise", *Eurospeech'97*, pp. 37-40, 1997.

[5] Cooke, M., Morris, A. and Green, P., "Missing data techniques for robust speech recognition", *ICASSP'97*, pp. 803-806, 1997.

[6] Drygajlo, A. and El-Maliki, M., "Speaker verification in noisy environment with combined spectral subtraction and missing data theory", *ICASSP'98*, pp. 121-124, 1998.

[7] Vizinho, A., Green, P., Cooke, M. and Josifovski, L., "Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: an integrated study", *Eurospeech'99*, pp. 2407-2410, 1999.

[8] Seltzer, M. L., Raj, B. and Stern, R. M., "Classifier-based mask estimate for missing feature method of robust speech recognition", *ICSLP'2000*, 2000.

[9] Ming, J. and Smith, F. J., "A probabilistic union model for sub-band based robust speech recognition", *ICASSP'2000*, pp. 1787-1790, 2000.

[10] Ming, J., Jancovic, P., Hanna, P., Stewart, D. and Smith, F.J., "Robust feature selction using probabilistic union models", *ICSLP'2000*, pp. 546-549, 2000.

[11] Rabiner, L. R., Wilpon, J. G. and Soong, F. K., "High performance connected digit recognition using hidden Markov models," *IEEE Trans. ASSP*, vol. 37, pp. 1214-1225, 1989.