

Robust Speaker Identification Using Posterior Union Models

Ji Ming[†], Darryl Stewart[†], Philip Hanna[†], Pat Corr[†],
Jack Smith[†], Saeed Vaseghi[‡]

[†]School of Computer Science, Queen's University Belfast, Belfast BT7 1NN, UK

[‡]Department of Electronic and Computer Engineering, Brunel University,
Middlesex UB8 3PH, UK

Abstract

This paper investigates the problem of speaker identification in noisy conditions, assuming that there is no prior knowledge about the noise. To confine the effect of the noise on recognition, we use a multi-stream approach to characterize the speech signal, assuming that while all of the feature streams may be affected by the noise, there may be some streams that are less severely affected and thus still provide useful information about the speaker. Recognition decisions are based on the feature streams that are uncontaminated or least contaminated, thereby reducing the effect of the noise on recognition. We introduce a novel statistical method, the posterior union model, for selecting reliable feature streams. An advantage of the union model is that knowledge of the structure of the noise is not needed, thereby providing robustness to time-varying unpredictable noise corruption. We have tested the new method on the TIMIT database with additive corruption from real-world nonstationary noise; the results obtained are encouraging.

1. Introduction

The process of speaker recognition is typically made more complex due to a number of factors. For example, an individual's physiological and emotional state will alter some characteristics of their voice. In addition, the mechanism that is used to collect the speech may also cause distortion, for example, different telephone handsets or microphones - known as *channel effects*. Potentially of even more significance are the effects of environmental noise, which can pollute the recorded voice sample. Current research has gone some way towards reducing speaker variability and channel effects. The most commonly used channel compensation techniques include feature compensation, model compensation and model adaptation. Methods for feature compensation make use of some form of linear or non-linear channel compensation (such as RASTA filtering, cepstral mean subtraction or artificial neural networks), that is applied to the acoustic analysis to produce features with improved robustness to channel effects (e.g. [1]–[5]). In addition to channel compensation in the feature domain, there have also been compensation techniques that are applied in the model and match score domains (e.g. [6]–[8]). The model adaptation techniques effectively use new data to learn channel characteristics or to keep the speaker's model up-to-date (e.g. [9][10]).

However, whilst there have been considerable studies on robustness against channel variability, there have been comparatively few studies on robustness against environmental noise [11]. The traditional noise-reduction techniques used in *speech* recognition may be applied to the problem of speaker recognition, as suggested in [12][13]. But these techniques usu-

ally assume a priori knowledge such as the spectral or cepstral characteristics of the noise. This knowledge may not be available in real-world applications involving unpredictable, abrupt noise (e.g. a door slam, a cough or a car horn), or noise with a time-varying nature (e.g. a passing car, a telephone ring or background music). The researchers in [14][15] have studied the use of the missing feature method for speaker recognition in noise, indicating that robust performance can be achieved by ignoring the part of the feature representation that is strongly distorted, and by basing the decision only on those parts of the feature representation that are uncontaminated or are least contaminated by the noise. The key problem is how to determine which part of the feature space is corrupt when no knowledge of the corruption is assumed. The missing feature method usually requires the noisy parts to be identified in order to remove the noise. The posterior union model is an alternative to the missing feature method which does not require knowledge concerning the identity of the noisy features, thereby providing robustness to unknown, time-varying noise corruption. In the following we first describe the multi-stream approach used to characterize the speech signal, and then we describe the posterior union model used for selecting the features streams for each utterance assuming no knowledge about the corrupting noise.

2. Speech Signal Characterization

A multi-stream approach is used to characterize the speech signal, and the feature streams with the highest SNRs are exploited for recognition. While a noise may affect all of the feature streams, there may be some streams that are more robust than others to a specific type of noise and therefore may still provide useful information for recognition. For example, the dynamic spectral features are usually more robust to channel distortion or slowly-varying noise than the static spectral features. Specifically, we divide the speech signal into multiple subbands, and calculate the static and dynamic feature vectors for each band independently of the other bands. This method has been used in speech recognition (e.g. [16][17][19]) for isolating local frequency corruption from spreading into the feature vectors of the other subbands. Since the speech in different frequency bands may have different local energies, it can be assumed that some of the bands will have a higher SNR than the others. These high SNR bands may contain more reliable information for correct recognition.

The TIMIT database was selected for the experiments. The database contains utterances from 630 speakers (438 male and 192 female), recorded under clean conditions with no intersession variability. In our experiments, the speech signal was divided into frames of 20 ms at a frame rate of 10 ms. For

each frame, we used a 12-channel mel-scale filter bank to estimate 12 log-amplitude spectral coefficients (i.e. log FB energies). These 12 log FB energies were decorrelated by using a decorrelation filter $H(z) = 1 - z^{-1}$, resulting in 12 decorrelated log FB energies. As indicated in [18], the decorrelated log FB energies may be used as an alternative to the conventional mel-frequency cepstral coefficients (MFCC) for robust speech recognition. The resulting 12 decorrelated log FB energies were then grouped uniformly into six subbands, with each subband containing two decorrelated log FB energies. For each decorrelated log FB energy, its first-order delta coefficient was also calculated, and the dynamic coefficients for all the 12 static coefficients were grouped into six subbands in the same way as for the static coefficients. For each subband, we thus have two feature streams: the static feature stream containing the two decorrelated log FB energies and the delta feature stream containing the two delta decorrelated log FB energies. For each frame with six subbands, we then have a total of 12 feature streams; each feature stream contains two coefficients, so the overall size of the feature vector for a frame is 24. As indicated in [20], the separation of the static and dynamic feature streams within a subband is important for reducing the effects of slowly-varying background or channel noise. These noises usually affect the static features more adversely than the dynamic features. By separating these streams, the dynamic features may be individually exploited to provide useful information about the speaker.

3. Posterior Union Model for Identification

The union model is a method for basing the recognition on the feature streams that are uncontaminated or least contaminated by the noise, thereby reducing the effect of the noise on recognition. To this end, it is similar to the missing feature method. However, the union model does not require the identification of the unreliable feature streams.

Let $X = (x_1, x_2, \dots, x_N)$ be a feature set consisting of N feature streams, to be classified into one of the K classes $\lambda_1, \lambda_2, \dots, \lambda_K$, representing K speakers. Assume that there are M ($0 \leq M < N$) streams in X being severely corrupted by noise, but neither the value of M nor the identity of the corrupted streams is known *a priori*. Denote by X_{N-M} the subset in X which contains the remaining $(N - M)$ reliable feature streams. The union model deals with the uncertainty of X_{N-M} by using the “or” (i.e. disjunction) operator to combine every $(N - M)$ sized subset of the streams, assuming that any one of the subsets may be X_{N-M} . Based on the probability theory for the union of random events, the conditional probability $P(X_{N-M} | \lambda_k)$ can be written as [19][20]

$$\begin{aligned} P(X_{N-M} | \lambda_k) &= P\left(\bigvee_{n_1 n_2 \dots n_{N-M}} x_{n_1} x_{n_2} \dots x_{n_{N-M}} | \lambda_k\right) \\ &\simeq \sum_{n_1 n_2 \dots n_{N-M}} P(x_{n_1} x_{n_2} \dots x_{n_{N-M}} | \lambda_k) \end{aligned} \quad (1)$$

where \bigvee denotes the “or” operator, $x_{n_1} x_{n_2} \dots x_{n_{N-M}}$ is a subset in X containing $(N - M)$ streams as a probable candidate for X_{N-M} , and the “or” operator (and hence the summation) is applied between all possible subsets of $(N - M)$ streams in X . Since (1) is the sum of the individual subset probabilities, its value is dominated by the subset probabilities with large values. Therefore, if we can assume that the “clean”-stream subset produces a large probability for the correct class (this should be achieved through training), then selecting the maximum value

of $P(X_{N-M} | \lambda_k)$ with respect to λ_k has a chance to get the correct class λ_k for X without requiring the identity of the M noisy streams. We call (1) the *conditional union model*, which has been applied previously to speech recognition for combining subband or segmental feature streams against band-limited or duration-limited noise (for a review of the model, see [21]). A disadvantage of this model is that, when the number of the noisy streams, i.e., M , is unknown, it is not possible to obtain an optimal estimate for M by maximizing $P(X_{N-M} | \lambda_k)$ with respect to M . This is because, for a specific λ_k , the values of $P(X_{N-M} | \lambda_k)$ for different M are of a different order of magnitude and are thus not directly comparable. This problem may be overcome by extending the union model from the conditional-probability formulation, i.e. (1), to a posterior-probability formulation. The *a posteriori union probability* of class λ_k given X_{N-M} is defined as

$$\begin{aligned} P(\lambda_k | X_{N-M}) &= \frac{P(X_{N-M} | \lambda_k)P(\lambda_k)}{P(X_{N-M})} \\ &= \frac{P(X_{N-M} | \lambda_k)P(\lambda_k)}{\sum_{j=1}^K P(X_{N-M} | \lambda_j)P(\lambda_j)} \end{aligned} \quad (2)$$

where $P(X_{N-M} | \lambda_k)$ is the conditional union probability as defined in (1) and $P(\lambda_k)$ is the prior probability for class λ_k . Substituting (1) into (2) and assuming an equal class prior, it can be shown that, similar to (1), (2) will be dominated by the subset conditional probabilities, i.e., $P(x_{n_1} x_{n_2} \dots x_{n_{N-M}} | \lambda_k)$, with large values. Therefore, if we assume that the clean subset produces a large conditional probability for the correct class, selecting the maximum $P(\lambda_k | X_{N-M})$ with respect to λ_k is likely to get the correct class λ_k for X without requiring the identity of the M noisy feature streams. The major difference between (2) and (1) is that (2) is normalized against M , such that the values of $P(\lambda_k | X_{N-M})$ for different M are of the same order of magnitude. An optimal classifier can thus be defined, which classifies an observation based on the maximum *a posteriori* (MAP) probability $P(\lambda_k | X_{N-M})$ with respect to both λ_k and M , i.e.,

$$X \in \lambda_k \quad \text{if} \quad \lambda_k = \arg \max_{\lambda_j} \max_M P(\lambda_j | X_{N-M}) \quad (3)$$

This classifier requires neither the identity nor the number of the noisy streams.

Assume that a test utterance is represented by a sequence of frame vectors $X_1^T = (X(1), X(2), \dots, X(T))$, where $X(t) = (x_1(t), x_2(t), \dots, x_N(t))$ is the frame feature vector at time t , consisting of N feature streams $x_n(t)$ as described in Section 2. This gives a 2-dimensional feature set $\{x_n(t) : n = 1, \dots, N; t = 1, \dots, T\}$. The above posterior union model can be applied to this feature set to select the feature streams leading to the MAP probability for recognition. In this study, we consider an one-dimensional case: the posterior union model is applied only across the stream indexes, not across the time indexes, i.e., no frame is skipped during the recognition. The *a posteriori* probability of λ_k given the frame sequence X_1^T can thus be written as

$$P(\lambda_k | X_1^T, M_1^T) = \prod_{t=1}^T P(\lambda_k | X_{N-M(t)}(t)) \quad (4)$$

where $P(\lambda_k | X_{N-M(t)}(t))$ is the *a posteriori* union probability of class λ_k and $X_{N-M(t)}$ denotes the subset in frame vector $X(t)$ containing $N - M(t)$ reliable streams, assuming that $X(t)$ contains N streams and $M(t)$ of which are corrupted; $M_1^T = (M(1), M(2), \dots, M(T))$ denotes the sequence

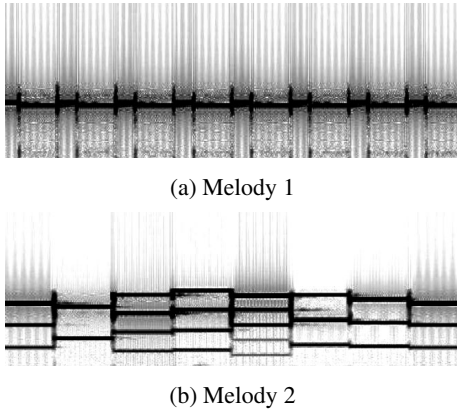


Figure 1: Spectra of mobile phone ring noises used in the experiments.

of $M(t)$ for the individual frame vectors in X_1^T . Applying (3) to (4) we thus have the recognition rule: $X_1^T \in \lambda_k$ if

$$\begin{aligned} \lambda_k &= \arg \max_{\lambda_j} \max_{M_1^T} P(\lambda_j | X_1^T, M_1^T) \\ &= \arg \max_{\lambda_j} \prod_{t=1}^T \max_{M(t)} P(\lambda_j | X_{N-M(t)}(t)) \end{aligned} \quad (5)$$

As indicated in (5), the value of $M(t)$ is optimized for every single frame, thereby providing robustness to time-varying noise corruption.

Since $P(\lambda_k | X_1^T, M_1^T)$ is a probability measure, its value can also be used as a confidence score to verify the identification result. Assume that, for a given observation X_1^T , the identified class is λ . The probability of this class equals $\max_{M_1^T} P(\lambda | X_1^T, M_1^T)$, as shown in (5). Verification can be conducted by comparing this probability (normalized by the length of the observation) with a pre-defined threshold.

4. Preliminary Results and Discussion

In the TIMIT database, each of the 630 speakers contributed 10 utterances, and each utterance has an average duration of about 3 seconds. As in [22], for each speaker, 8 utterances were used to train a speaker model and the remaining 2 utterances were used for testing. This gives a total of 1260 test utterances across all the 630 speakers. The multi-stream feature format described in Section 2 was used in the posterior union model. We assumed independence between the feature streams. For every speaker, each feature stream was modeled using a Gaussian mixture model (GMM) with 32 mixtures. For comparison, we also implemented a baseline recognition system based on GMM, using a feature vector of the same size (i.e., 24, 12 MFCC plus 12 delta MFCC) for each frame and also 32 mixtures for each speaker. Both the union model and the baseline model were trained using clean speech data.

Two mobile phone ring noises, labelled as *melody 1* and *melody 2*, were used to corrupt the test utterances. As shown in Fig. 1, both noises exhibit a time-varying nature, especially for melody 2. These noises were added, respectively, to each of the test utterances to simulate real-world time-varying noise corruption. Fig. 2 shows examples of the noisy speech utterances used in the recognition. Due to the difficulty in estimating the structures of the nonstationary noise, no noise-reduction techniques

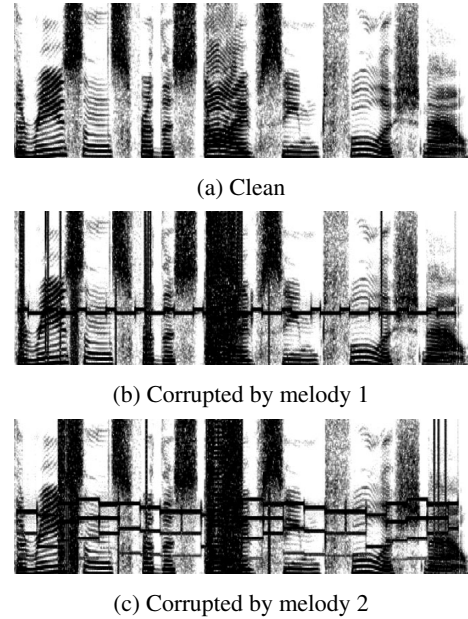


Figure 2: Spectra of clean and noisy speech utterances.

were implemented in the union model and baseline model.

Table 1 shows the speaker identification results using clean speech utterances, obtained by the baseline model and the new posterior union model. The two models achieved similar accuracy in clean speech conditions. Tables 2 and 3 show the identification accuracy obtained by the two models for the melody 1 and melody 2 noise conditions, respectively, as a function of the SNR. As shown in these two tables, the baseline model was sensitive to the background noise and its performance significantly degraded under both noisy conditions. The noises also caused problems to the posterior union model, but the model showed much stronger robustness in comparison to the baseline model.

If verification is introduced, it would be expected for the identification accuracy to improve when raising the rejection rate. Fig. 3 shows the results. For example, at a rejection rate of 10%, it is possible to improve the identification accuracy from 78.0% to 83.4% for the melody 1 noise, and from 67.5% to 72.8% for the melody 2 noise, both with an SNR=10 dB. Further improvement would be expected by applying the union model to the 2-dimensional (i.e. time-frequency) feature space, not only selecting the least-contaminated feature streams within each frame, but also selecting the least-contaminated frames.

5. References

- [1] Reynolds, D. A., "Experimental evaluation of features for robust speaker identification," IEEE Trans. Speech and Audio Processing, Vol. 2, pp. 639-643, 1994.
- [2] Mammon, R., Zhang, X. and Ramachandran, R. P., "Robust speaker recognition - a feature-based approach," IEEE Signal Processing Magazine, pp. 58-71, Spet. 1996.
- [3] Reynolds, D. A., "The effects of handset variability on spaker recognition performance: experiments on the Switchboard corpus," ICASSP'96, pp. 113-116, 1996.
- [4] van Vaaren, S., "Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch," ICSLP'96, pp. 1788-1791, 1996.

- [5] Heck, L. P., Konig, Y., Sonmez, M. K. and Weintraub, M., "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," *Speech Communication*, Vol. 31, pp. 181-192, 2000.
- [6] Gish, H. and Schmidt, M., "Text-independent speaker identification," *IEEE Signal Processing Magazine*, pp. 18-32, Oct. 1994.
- [7] Murthy, H. A., Beaufays, F., Heck, L. P. and Weintraub, M., "Robust text-independent speaker identification over telephone channels," *IEEE Trans. Speech and Audio Processing*, Vol. 7, pp. 554-568, 1999.
- [8] Teunen, R., Shahshahani, B. and Heck, L., "A model-based transformational approach to robust speaker recognition," *ICSLP'2000*, 2000.
- [9] Legetter, J. C. and Woodland, P. C., "Flexible speaker adaptation using maximum likelihood linear regression," *Eurospeech'95*, pp. 1155-1158, 1995.
- [10] Lamel, L. F. and Gauvain, J. L., "Speaker verification over the telephone," *Speech Communication*, Vol. 31, pp. 141-154, 2000.
- [11] Doddington, G. R. et al., "The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective", *Speech Communication*, Vol. 31, pp. 225-254, 2000.
- [12] Matsui, T., Kanno, T. and Furui, S., "Speaker recognition using HMM composition in noisy environments," *Computer Speech and Language*, Vol. 10, pp. 107-116, 1996.
- [13] Ortega-Garcia, J. and Gonzalez-Rodriguez, J., "Overview of speaker enhancement techniques for automatic speaker recognition," *ICSLP'96*, pp. 929-932, 1996.
- [14] 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.
- [15] Besacier, L., Bonastre, J. F. and Fredouille, C., "Localization and selection of speaker-specific information with statistical modelling", *Speech Communication*, Vol. 31, pp. 89-106, 2000.
- [16] Bourlard, H. and Dupont, S., "A new ASR approach based on independent processing and recombination of partial frequency bands", *ICSLP'96*, pp. 426-429, 1996.
- [17] Hermansky, H., Tibrewala, S. and Pavel, M., "Towards ASR on partially corrupted speech", *ICSLP'96*, pp. 462-465, 1996.
- [18] Paliwal, K. K., "Decorrelated and lifted filter-bank energies for robust speech recognition," *Eurospeech'99*, pp. 85-88, 1999.
- [19] Ming, J. and Smith, F. J., "Union: a new approach for combining subband observations for noisy speech recognition," *Speech Communication*, Vol. 34, pp. 41-55, 2001.
- [20] Ming, J., Jancovic, P. and Smith, F. J., "Robust speech recognition using probabilistic union models," *IEEE Trans. Speech Audio Processing*, Vol. 10, pp.403-414, 2002.
- [21] Ming, J. and Smith, F. J., "Speech recognition with unknown partial feature corruption - a review of the union model," to appear in *Computer Speech and Language*.
- [22] Reynolds, R. A., "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, Vol. 17, pp. 91-108, 1995.

Table 1: *Speaker identification accuracy (%) using clean speech utterances, for the posterior union model, compared to a baseline model*

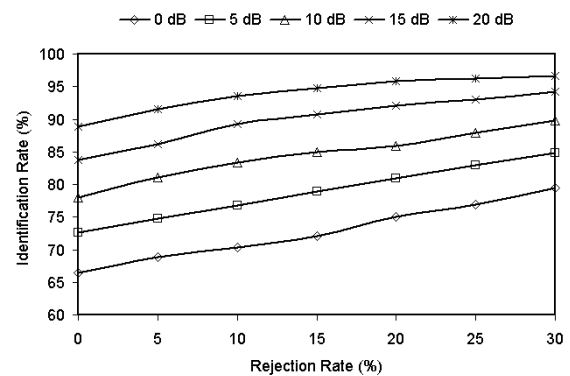
	Posterior union	Baseline
	97.4	97.6

Table 2: *Speaker identification accuracy (%) in melody 1 noisy conditions.*

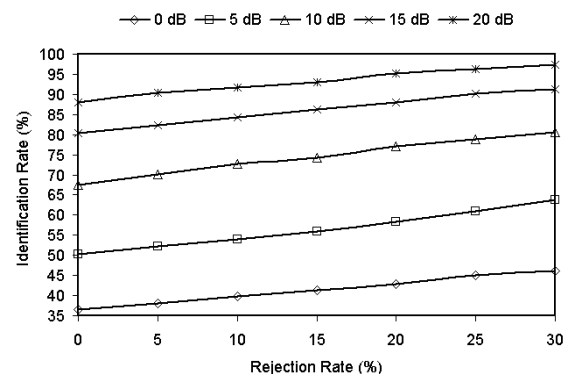
SNR (dB)	Posterior union	Baseline
20	88.8	61.4
15	83.8	39.3
10	78.0	22.9
5	72.6	11.8
0	66.5	6.4

Table 3: *Speaker identification accuracy (%) in melody 2 noisy conditions.*

SNR (dB)	Posterior union	Baseline
20	88.1	76.8
15	80.4	52.8
10	67.5	29.7
5	50.2	12.5
0	36.6	4.8



(a) Melody 1 noisy conditions



(b) Melody 2 noisy conditions

Figure 3: *Identification rate as a function of rejection rate, in melody 1 and melody 2 noisy conditions with different SNR.*