Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 75390, Pages 1–12
DOI 10.1155/ASP/2006/75390

A Posterior Union Model with Applications to Robust Speech and Speaker Recognition

Ji Ming,[1] Jie Lin,[2] and F. Jack Smith[1]

[1] *School of Computer Science, Queen's University Belfast, Belfast BT7 1NN, UK*
[2] *School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China*

Received 13 January 2005; Revised 12 December 2005; Accepted 14 December 2005

Recommended for Publication by Douglas O'Shaughnessy

This paper investigates speech and speaker recognition involving partial feature corruption, assuming unknown, time-varying noise characteristics. The probabilistic union model is extended from a conditional-probability formulation to a posterior-probability formulation as an improved solution to the problem. The new formulation allows the order of the model to be optimized for every single frame, thereby enhancing the capability of the model for dealing with nonstationary noise corruption. The new formulation also allows the model to be readily incorporated into a Gaussian mixture model (GMM) for speaker recognition. Experiments have been conducted on two databases: TIDIGITS and SPIDRE, for speech recognition and speaker identification. Both databases are subject to unknown, time-varying band-selective corruption. The results have demonstrated the improved robustness for the new model.

1. INTRODUCTION

Speech and speaker recognition systems need to be robust against unknown partial corruption of the acoustic features, where some of the feature components may be corrupted by noise, but knowledge about the corruption, including the number and identities of the corrupted components and the characteristics of the corrupting noise, is not available. This problem has been addressed recently by the missing-feature methods (see, e.g., [1–10]), which have focused on how to identify and thereby remove those feature components that are severely distorted by noise and thus provide unreliable information for recognition. A number of methods have been suggested for identifying the corrupt data, for example, based on a measurement of the local signal-to-noise ratio (SNR) or other noise characteristics such as the statistical distribution [3–5, 10], based on knowledge of the speech such as the harmonic structure of voiced speech [7], and based on a combination of auditory scene analysis and SNR for mixed voiced and unvoiced speech [8]. A more recent development, termed fragment decoder, is detailed in [11]. The fragment decoder models an utterance as fragments (time-frequency regions) of speech and background. The missing-feature theory is incorporated into the model to facilitate the search for the most likely speech fragments forming the speech utterance. In this paper, we describe an alternative, the posterior union model, as a complement to the above methods. The posterior union model is an extension of our previous conditional-probability union model described in [12, 13]. The aims of the extension are two folds: (1) enhancing the model's capability for dealing with nonstationary noise corruption, and (2) enabling the incorporation of the model into Gaussian mixture model (GMM) based speaker recognition.

As an alternative to the missing-feature methods, the union model aims to lift the requirement for identifying the noisy features. Assume a feature set comprising N components, M of which are corrupt, and recognition is ideally based only on the remaining $(N-M)$ clean components. The union model deals with the uncertainty of the clean components by forming a union of all possible combinations of $(N-M)$ components, which therefore includes the combination of the $(N-M)$ clean components, and by assuming that the probability of the union will be dominated by this all-clean component combination for correct recognition. This effectively reduces the problem of identifying the noisy components to a problem of estimating the number of the noisy components, that is, M, required to form the union. We term this number *the order of the union model*.

Previously we have studied the formulation of the union model using the conditional probabilities of the features, and applied the model to subband-based speech recognition

[12, 13]. In those systems, each speech frame is modeled by a feature vector consisting of short-time subband spectral measurements. A major drawback of this conditional-probability model is the lack of effective means for estimating the order, that is, the number of corrupted subbands within each frame. Towards a solution, a heuristic method was suggested in [14], assuming the use of a multistate hidden Markov model (HMM) for modeling a speech utterance. The method compares the state occupancies associated with each hypothesized order with the state occupancies for clean training utterances, and assumes that the model with the correct order should produce a state-occupancy distribution similar to the state-occupancy distribution for the clean utterances due to the isolation of noisy subbands. In estimating the state occupancies for a test utterance, the method assumes the same number of noisy subbands (i.e., order) throughout the utterance. This method thus offers only a suboptimal performance in nonstationary noise conditions, in which different frames may involve different subband corruption due to the time-varying nature of the noise. Moreover, this state-occupancy method becomes invalid for an HMM with only a single state, for example, a GMM. GMMs are commonly used for modeling speakers for speaker identification and verification (e.g., [15]).

In this paper, we describe an extension of the union model from the conditional-probability formulation to a posterior-probability formulation, as a solution to the above problem. The new formulation allows the order to be optimized for every single frame subject to an optimality criterion, to enhance the capability of the model for dealing with nonstationary noise corruption. The frame-by-frame order estimation also enables the incorporation of the model into GMM-based speaker recognition systems, to provide robustness to unknown, time-varying partial feature corruption.

The remainder of this paper is organized as follows. Section 2 formulates the problem. Section 3 describes the new posterior union model and its incorporation into the HMM/GMM framework for speech and speaker recognition. The experimental results are presented in Section 4, followed by a conclusion in Section 5.

2. PROBLEM FORMULATION

Assume a feature set $X = (x_1, x_2, \ldots, x_N)$ consisting of N components, where x_n represents the nth component, to be classified into one of the K classes, $C_1, C_2, \ldots, C_K$. In speech recognition, for example, X may be a frame feature vector consisting of N feature streams, and C_k corresponds to the underlying speech state forming a phone or a word. Assume that within the N components there are M components being corrupted, and further assume that the corruption is partial, that is, $0 \le M < N$ ($M = 0$ means no corruption). To reduce the effect of the noise, classification can be based on the marginal probability of the remaining $(N - M)$ clean components, with the noisy components being removed to improve mismatch robustness (the missing-feature theory). Without knowledge of the identity of the noisy components, these $(N - M)$ clean components could be any one of the combinations of $(N - M)$ components taken from X. Therefore the random nature of the clean components can be modeled by the union of all these combinations. Use a simple case as an example, in which X is a 3-component feature set $X = (x_1, x_2, x_3)$ and there is one component (say x_1) that is noisy but the identity of the noisy component is not known. Consider the union of all possible combinations of two components. Denoting the union variable by χ_2, $\chi_2 = x_1x_2 \vee x_1x_3 \vee x_2x_3$, where $\vee$ stands for the disjunction (i.e., "or") operator. The union includes the true clean combination (x_2x_3) that contains all the clean components and no others, and the noisy combinations (x_1x_2, x_1x_3) that are affected by the noisy component x_1. Consider the probability of the union χ_2 associated with class C_k, $P(\chi_2 \mid C_k)$. This can be written as

$$\begin{aligned} P(\chi_2 \mid C_k) &= \frac{P(\{x_1x_2 \wedge C_k\} \vee \{x_1x_3 \wedge C_k\} \vee \{x_2x_3 \wedge C_k\})}{P(C_k)} \\ &= P(x_1x_2 \mid C_k) + P(x_1x_3 \mid C_k) + P(x_2x_3 \mid C_k) \\ &\quad - P(x_1x_2 \wedge x_1x_3 \mid C_k) - P(x_1x_2 \wedge x_2x_3 \mid C_k) \\ &\quad - P(x_1x_3 \wedge x_2x_3 \mid C_k) \\ &\quad + P(x_1x_2 \wedge x_1x_3 \wedge x_2x_3 \mid C_k) \\ &= P(x_1x_2 \mid C_k) + P(x_1x_3 \mid C_k) + P(x_2x_3 \mid C_k) \\ &\quad + \rho(x_1x_2, x_1x_3, x_2x_3), \end{aligned} \tag{1}$$

where $\wedge$ is short for the "and" operator, and the last term $\rho(x_1x_2, x_1x_3, x_2x_3)$ summarizes the joint probabilities between and across the combinations x_1x_2, x_1x_3, and x_2x_3 included as a result of the probability normalization. Equation (1) includes all marginal probabilities of two components, and hence includes $P(x_2x_3 \mid C_k)$ of the two clean components, that is, the marginal probability sought for recognition. In our previous speech recognition experiments based on subband features (e.g., [12]), the joint probabilities between and across the combinations, $\rho(\cdot)$, were found to be unimportant in the sense that they were smaller than the corresponding marginal probabilities (e.g., $P(x_1x_2 \wedge x_1x_3 \mid C_k) \le P(x_1x_2 \mid C_k)$). Additionally, $\rho(\cdot)$ is affected by noise (x_1 in the above example), which reduces the value of $\rho(\cdot)$ for the correct class to be recognized. Therefore for maximum probability-based recognition applications, $\rho(\cdot)$ may be ignored in the computation. Ignoring $\rho(\cdot)$, (1) is a sum of the marginal probabilities of two components and is dominated by the probabilities with large values. Assume that the observation probability distribution $P(\cdot \mid C_k)$ for each class C_k is trained using clean data, such that the probability for the occurrence of clean data is maximized (e.g., the maximum likelihood criterion). Then (1) should reach a high value for the correct class C_k due to the maximization of $P(x_2x_3 \mid C_k)$ for the class given the clean feature components x_2x_3. For an incorrect class C_k, the value of $P(x_2x_3 \mid C_k)$ should be low because of the mismatch between the clean test data x_2x_3 and

the wrong class model $P(\cdot \mid C_k)$. In other words, given no information about the identity of the noisy component, we may use the union probability $P(\chi_2 \mid C_k)$ as an approximation for the marginal probability of the true clean components $P(x_2x_3 \mid C_k)$, in the sense that both produce large values for the correct class. In the above example we assume that the noisy component is x_1, but the same observation applies to the cases in which the noisy component is x_2 or x_3.

The above example can be extended to a general N-component feature set $X = (x_1, x_2, \ldots, x_N)$, assuming M unknown noisy components and hence $(N-M)$ unknown clean components. Denote by χ_{N-M} the union of all possible combinations of $(N - M)$ components. The probability of the union given class C_k, ignoring the joint probabilities between and across the combinations (i.e., $\rho(\cdot)$), can be written as

$$P(\chi_{N-M} \mid C_k) = P\left(\bigvee_{n_1 n_2 \cdots n_{N-M}} x_{n_1} x_{n_2} \cdots x_{n_{N-M}} \mid C_k \right) \propto \sum_{n_1 n_2 \cdots n_{N-M}} P(x_{n_1} x_{n_2} \cdots x_{n_{N-M}} \mid C_k), \tag{2}$$

where $x_{n_1} x_{n_2} \cdots x_{n_{N-M}}$ is a combination in X consisting of $(N - M)$ components, with the indices $n_1 n_2 \cdots n_{N-M}$ representing a combination of $\{1, 2, \ldots, N\}$ taking $(N - M)$ at a time, and the "or" and the subsequent summation are taken over all possible such combinations. As described above, given no knowledge of the identity of the M noisy components, $P(\chi_{N-M} \mid C_k)$ defined in (2) can be used as an approximation for the marginal probability of the $(N - M)$ clean components, which is included in the sum, for maximum probability-based recognition of the correct class. The proportionality in (2) is due to the omission of $\rho(\cdot)$. Note that (2) is not a function of the identity of the clean components but only a function of the size of the clean components, determined by the number of noisy components M. We therefore effectively turn the problem of identifying the noisy components to a problem of estimating the number of the noisy components required to form the union. We call M *the order of the union model*. Estimating M without assuming knowledge of the noise is the focus of the paper. In implementation, we assume independence between the individual feature components. So $P(\chi_{N-M} \mid C_k)$ can be written as

$$P(\chi_{N-M} \mid C_k) \propto \sum_{n_1 n_2 \cdots n_{N-M}} P(x_{n_1} \mid C_k) P(x_{n_2} \mid C_k) \cdots P(x_{n_{N-M}} \mid C_k), \tag{3}$$

where $P(x_n \mid C_k)$ is the probability of feature component x_n given class C_k.

We particularly call the above model, (2) and (3), the *conditional union model of order M* as they model the conditional probability of the observation (feature set) associated with each class. The model may be used to accommodate M corrupted feature components, within N given feature components, without requiring the identity of the noisy components. However, given no knowledge about the noise, estimating M (i.e., the order) itself can be a difficult task with the conditional union model. Equation (3) suggests that it is not possible to obtain an optimal estimate for M by maximizing $P(\chi_{N-M} \mid C_k)$ with respect to M. This is because, for a specific C_k, the values of $P(\chi_{N-M} \mid C_k)$ for different M are of a different order of magnitude and thus not directly comparable.[1] In this paper we present a new formulation, namely, the posterior-probability formulation, for the union model to overcome this problem.

3. THE POSTERIOR UNION MODEL

3.1. The model

Using the same notation as above, let $X = (x_1, x_2, \ldots, x_N)$ be a feature set with N components, to be classified into one of the K classes $C_1, C_2, \ldots, C_K$. Assume that there are M $(0 \leq M < N)$ components in X being corrupted, but neither the value of M nor the identity of the corrupted components is known a priori. Use the union χ_{N-M} defined above to model the $(N - M)$ unknown clean components. The classification can be performed based on the a posteriori union probability $P(C_k \mid \chi_{N-M})$ of class C_k given χ_{N-M}, which is defined by

$$P(C_k \mid \chi_{N-M}) = \frac{P(\chi_{N-M} \mid C_k) P(C_k)}{\sum_{j=1}^{K} P(\chi_{N-M} \mid C_j) P(C_j)}, \tag{4}$$

where $P(\chi_{N-M} \mid C_k)$ is the conditional union probability of order M and $P(C_k)$ is the prior probability of class C_k, which is assumed not to be a function of the order M. Substituting (3) into (4) for $P(\chi_{N-M} \mid C_k)$, we can have

$$P(C_k|\chi_{N-M}) \propto \frac{\sum_{n_1 n_2 \cdots n_{N-M}} P(x_{n_1}|C_k) P(x_{n_2}|C_k) \cdots P(x_{n_{N-M}}|C_k) \cdot P(C_k)}{P(\chi_{N-M})}, \tag{5}$$

where by definition, $P(\chi_{N-M})$ is given by

$$P(\chi_{N-M}) = \sum_{j=1}^{K} \left[\sum_{n_1 n_2 \cdots n_{N-M}} P(x_{n_1} \mid C_j) P(x_{n_2} \mid C_j) \cdots P(x_{n_{N-M}} \mid C_j) \right] \times P(C_j). \tag{6}$$

Since $P(\chi_{N-M})$ is not a function of the class index and the identity of the clean components (but only a function of the size of the clean components), the comparison of $P(C_k \mid \chi_{N-M})$ is decided by the numerator, which is a sum as shown in (5) and thus dominated by the marginal conditional probabilities $P(x_{n_1} \mid C_k) P(x_{n_2} \mid C_k) \cdots P(x_{n_{N-M}} \mid C_k)$ with large

[1] For example, assume a 3-component feature set $X = (x_1, x_2, x_3)$. Comparing the conditional union probabilities of orders 1 and 2 leads to the comparison between the value of $P(x_1)P(x_2) + P(x_1)P(x_3) + P(x_2)P(x_3)$ and the value of $P(x_1) + P(x_2) + P(x_3)$ (the condition C_k is omitted in these probabilities for clarity). The comparison may always favor the latter assuming that $P(x_1)$, $P(x_2)$, and $P(x_3)$ are all within the range of $[0, 1]$.

values. Therefore, as for the conditional union model (3), if we assume that the clean components produce a large marginal conditional probability for the correct class, then selecting the maximum posterior union probability $P(C_k \mid \chi_{N-M})$ with respect to C_k is likely to obtain the correct class without requiring the identity of the M noisy components. A major difference between (3) and (5) is that the posterior union probability is normalized for the number of the clean components, or equivalently the order M, always producing a value in the range $[0, 1]$ for any value of M within the range $0 \le M < N$. This makes it possible to compare the probabilities associated with different M and to obtain an estimate for M based on the comparison. Specifically, for each class C_k, we can obtain an estimate for M by maximizing the posterior union probability $P(C_k \mid \chi_{N-M})$ of the class, that is,

$$\widehat{M} = \arg\max_{M} P(C_k \mid \chi_{N-M}), \tag{7}$$

where $\widehat{M}$ represents the estimate of M. An insight into decision (7) may be obtained by rewriting (4) in terms of the likelihood ratios between the classes. Dividing both the numerator and denominator of (4) by $P(\chi_{N-M} \mid C_k)$ gives

$$P(C_k \mid \chi_{N-M}) = \frac{P(C_k)}{P(C_k) + \sum_{j\neq k}^{K} P(C_j)\left(P(\chi_{N-M} \mid C_j)/P(\chi_{N-M} \mid C_k)\right)}. \tag{8}$$

Therefore, maximizing $P(C_k \mid \chi_{N-M})$ for M is equivalent to maximizing the likelihood ratios $P(\chi_{N-M} \mid C_k)/P(\chi_{N-M} \mid C_j)$ for C_k compared to all $C_j \neq C_k$. For C_k being the correct class, this estimate for M tends to be an optimal estimate since only the clean feature combination, containing the maximum number of clean components, is most likely to produce maximum likelihood ratios between the correct and incorrect classes. For C_k being an incorrect class, (7) will also lead to an M for a feature combination, likely including some noisy feature components, which favors C_k. Robustness is expected if this effect is outweighed by the maximization of the likelihood for the correct class due to the selection of clean or least-distorted feature components.

We call $P(C_k \mid \chi_{N-M})$ the *posterior union probability of order M*. The new model improves over the conditional union model by retaining the advantage of requiring no identity of the noisy components, and by additionally providing a means of estimating the model order, that is, the number of noisy components, through maximizing the class posterior (i.e., (7)). In the following we describe the incorporation of the new model into an HMM/GMM for subband-based speech and speaker recognition, assuming that speech signals are subject to band-selective corruption, but knowledge about the identity and the number of the noisy subbands is not available.

3.2. Incorporation into HMM/GMM

The above posterior union model can be incorporated into an HMM for modeling frame-level subband features subject to unknown band-selective corruption. The system uses $P(C_k \mid \chi_{N-M})$ for the state emission probability, with C_k corresponding to a state, X corresponding to a frame vector comprising N short-time subband components, and χ_{N-M} modeling the clean subband components in the frame, of an unknown order M. Following (4), the posterior union probability of state s given frame vector X can be written as

$$P(s \mid \chi_{N-M}) = \frac{P(\chi_{N-M} \mid s)P(s)}{\sum_{s'} P(\chi_{N-M} \mid s')P(s')}, \tag{9}$$

where $P(s)$ is a state prior, $P(\chi_{N-M} \mid s)$ is the conditional union probability in state s which is approximated by (3) with C_k replaced by s (assuming independence between the subbands), that is,

$$P(\chi_{N-M} \mid s) \propto \sum_{n_1 n_2 \cdots n_{N-M}} P(x_{n_1} \mid s)P(x_{n_2} \mid s) \cdots P(x_{n_{N-M}} \mid s), \tag{10}$$

where $P(x_n \mid s)$ is the state emission probability for subband component x_n. The summation in the denominator of (9) is over all possible states for the frame. To incorporate (9) into an HMM, we first express the traditional HMM in terms of the posterior probabilities of the states. Denote by $X_1^T = (X(1), X(2), \ldots, X(T))$ a speech utterance of T frames, where $X(t)$ is the frame vector at time t, and by $S_1^T = (s_1, s_2, \ldots, s_T)$ the state sequence for X_1^T. The joint probability of X_1^T and S_1^T based on an HMM with parameter set λ is defined as

$$\begin{aligned} P(X_1^T, S_1^T \mid \lambda) &= \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} P(X(t) \mid s_t) \\ &= \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \frac{P(X(t) \mid s_t)}{P(X(t))} P(X(t)) \\ &= \pi_{s_0} \left[\prod_{t=1}^{T} \frac{a_{s_{t-1}s_t}}{P(s_t)} P(s_t \mid X(t)) \right] \left[\prod_{t=1}^{T} P(X(t)) \right], \end{aligned} \tag{11}$$

where $P(s_t \mid X(t))$ is the posterior probability of state s_t given frame $X(t)$, $P(s_t)$ is the state prior, and $[\pi_i]$ and $[a_{ij}]$ are the initial state and state transition probabilities, respectively. The last product, $\prod_{t=1}^{T} P(X(t))$, is not a function of the state index and thus has no effect in recognition. Equation (11) may be further simplified by assuming an equal state prior probability $P(s_t)$.[2] Substituting (9) into (11) for each $P(s_t \mid X(t))$, with the optimization over the order (i.e., (7)) included and the time index indicated, we obtain a new HMM for recognition:

$$P(X_1^T, S_1^T \mid \lambda) \propto \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \max_{M_t} P(s_t \mid \chi_{N-M_t}(t)), \tag{12}$$

[2] Alternatively, $P(s_t)$ may be derived from $[\pi_i]$ and $[a_{ij}]$ based on the Markovian state assumption. But this did not turn out to perform better than the simple uniform assumption for $P(s_t)$ as experienced in our experiments.

where M_t represents the order (i.e., the number of corrupted subbands) in frame $X(t)$. Equation (12) can be implemented using the conventional Viterbi algorithm, with an additional maximization for estimating the order for each frame. This frame-by-frame order estimation enhances the capability of the model for dealing with nonstationary band-selective noise that affects different numbers of subbands at different frames.

The above model can be modified for speaker identification. Assume that each speaker is modeled by a single-state HMM, with the state emission probability modeled by a GMM. Given an utterance with T frames X_1^T, the union-based probability for speaker γ can be written, based on (12), as

$$P(X_1^T \mid \gamma) \propto \prod_{t=1}^{T} \max_{M_t} P(\gamma \mid \chi_{N-M_t}(t)), \tag{13}$$

where $P(\gamma \mid \chi_{N-M})$ is the posterior union probability of speaker γ given frame X, defined below

$$P(\gamma \mid \chi_{N-M}) = \frac{P(\chi_{N-M} \mid \gamma)P(\gamma)}{\sum_{\gamma'} P(\chi_{N-M} \mid \gamma')P(\gamma')}, \tag{14}$$

where $P(\gamma)$ is the prior probability for speaker γ, and $P(\chi_{N-M} \mid \gamma)$ is the conditional union probability of frame X given speaker γ, which is approximated by (3) with C_k replaced by the speaker index. The summation in the denominator of (14) is taken over all speakers in consideration. As shown in (13), the maximization over the order is performed on a frame-by-frame basis, as in the multistate HMM (12) for speech recognition. In our implementation, the conditional probability of a frame X, that is, $P(X \mid C_k)$, where X is a N-component feature vector and C_k can be a state or speaker index, is modeled by using a GMM. The conditional union probability (3), of order M, is obtained from $P(X \mid C_k)$ by combining all the marginal versions of $P(X \mid C_k)$ with $(N - M)$ components.

4. EXPERIMENTAL RESULTS

4.1. Experiments on TIDIGITS for speech recognition

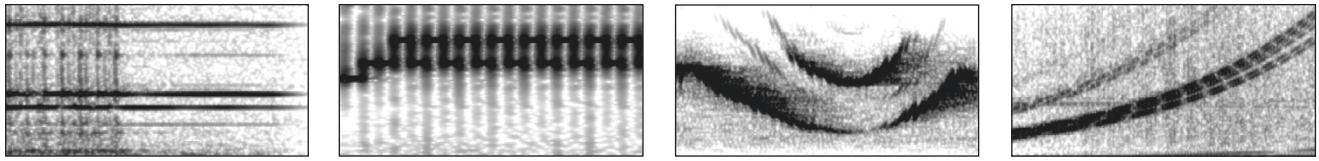
The above model (12) based on subband features has been tested for speech recognition involving unknown, time-varying band-selective corruption. The TIDIGITS database [16] was used in the experiments. The database contains utterances from 225 adult speakers, divided into training and testing sets, for speaker-independent connected digit recognition. The test set provided 6196 utterances from 113 speakers. The number of digits in the test utterances may be two, three, four, five, or seven, each roughly of an equal number of occurrences, and we assumed no advance knowledge of the number of digits in a test utterance.

Each speech frame was modeled by a feature vector consisting of components from individual subbands. Two different methods have been used to create the subband features. The first method produces the subband MFCC (mel-frequency cepstral coefficients) [12, 13], obtained by first grouping the mel-scale filter bank uniformly into subbands, and then performing a separate DCT within each subband to obtain the MFCC for that subband. It is assumed that the separation of the DCT among the subbands helps to prevent the effect of a band-selective noise from being spread over the entire feature vector, as usually occurs within the traditional full-band MFCC. The second method derives the subband features from the decorrelated log filter-bank amplitudes, obtained by filtering the amplitudes using a high-pass filter (more details will be described later). Our experiments for both speech recognition and speaker identification indicate that the two methods are equally effective for dealing with band-selective corruption. Article [12] described the use of the subband MFCC for speech recognition over the TIDIGITS database, based on the conditional union model that uses (10) as the state emission probability. To decide the model order M (i.e., the number of noisy subbands), the model assumes that the correct order, which correctly isolates the noisy bands from the clean bands, will result in a state-occupancy pattern that closely matches the state-occupancy pattern shown by the clean utterances [14]. However, for an utterance with T frames and N subbands, there could be N^T different order combinations and thus potentially N^T different state-occupancy patterns. To make the search for the best state-occupancy pattern/order computationally tractable, the model assumes that the order remains invariant within an utterance and changes only from utterance to utterance. This reduces the number of searches for each test utterance to N but compromises the ability of the model for dealing with nonstationary noise that affects a varying number of subbands over the duration of an utterance. The focus of this subsection is to compare this conditional union model, described above and detailed in [12–14], with the new posterior union model that uses (9) as the state emission probability and estimates the order on a frame-by-frame basis as shown in (12). For this comparison, the same feature format and the same test conditions as in [12] are implemented for the new posterior union model, such that any observed improvement in recognition performance would be mainly attributable to the improved estimation for the order in the new posterior union model. The effectiveness of the subband features derived from the decorrelated log filter-bank amplitudes is demonstrated through experiments for speaker identification, described in the next subsection.

The speech was divided into frames of 256 samples at a frame period of 128 samples. For each frame, a 30-channel mel-scale filter bank was used to obtain 30 log filter-bank amplitudes. These were uniformly grouped into five subbands. For each subband, three MFCC and three delta MFCC, obtained over a window of ± 2 frames within the same subband, were derived as the feature components for the subband. Thus, for this 5-band system, there was a feature vector of ten streams for each frame:

$$X(t) = (x_1(t), \ldots, x_5(t), \Delta x_1(t), \ldots, \Delta x_5(t)), \tag{15}$$

where $x_n(t)$ and $\Delta x_n(t)$, each being a vector of three elements,



(a) Telephone ring (b) Whistle (c) Contact (d) Connect

Figure 1: Spectra of the real-world noise data used in speech recognition experiments.

Table 1: Digit string accuracy (%) in nonstationary real-world noise, for the posterior union model, compared with the conditional union model, the product model, and the baseline full-band HMM.

SNR (dB)	Noise type	Posterior union	Conditional union	Product model	Baseline full-band
Clean		96.42	96.21	96.48	97.53
20	Ring	92.03	91.69	87.36	93.59
	Whistle	93.88	93.29	87.17	88.36
	Contact	93.46	91.80	79.41	89.33
	Connect	91.74	91.14	76.19	89.36
15	Ring	89.30	87.90	73.55	83.44
	Whistle	93.22	92.64	77.02	74.31
	Contact	92.79	90.43	63.02	76.39
	Connect	88.02	87.04	52.05	72.39
10	Ring	85.73	81.99	49.79	60.23
	Whistle	92.82	90.95	62.62	50.44
	Contact	91.56	88.15	41.98	53.62
	Connect	81.71	79.21	24.44	41.59
5	Ring	76.78	73.87	28.18	34.49
	Whistle	90.57	88.75	46.29	25.87
	Contact	88.56	85.31	24.27	30.57
	Connect	68.62	65.80	8.86	16.03
0	Ring	64.93	62.90	14.61	17.75
	Whistle	86.60	84.88	31.00	8.28
	Contact	84.31	81.81	12.86	14.27
	Connect	48.31	44.96	2.68	4.50

represent the static and delta MFCC for the nth subband, respectively. This frame vector was modeled by the posterior union model (9) and the conditional union model (10), with $N = 10$ and an order range $0 \leq M_t \leq 5$, allowing from no feature stream corruption up to five feature stream corruption within each frame. In addition to the two union models, the results produced by two other models are also included. The first is a "product" model, which uses the same subband features as the union model but ignores no subband from the computation, which is therefore equivalent to the conditional union model with order $M = 0$ ((10), which is reduced to a product of the probabilities of the individual subband streams when $M = 0$). The second is a baseline full-band HMM, based on full-band features for each frame (10 MFCC and 10 delta MFCC, derived from a mel-scale filter bank with 20 channels). All the models have the same HMM topology: each digit was modeled by a left-to-right HMM with ten states, and each state consisted of eight Gaussian mixtures with diagonal covariance matrices.

Figure 1 shows the real-world noises used in the test, including a telephone ring, a whistle, and the sounds of "contact" and "connect," extracted from an Internet tool. These noises each had a dominant band-selective nature, and the noises "contact" and "connect" were particularly nonstationary. These noises were added, respectively, to each of the test utterances with different levels of SNR. Table 1 presents the

digit string accuracy[3] obtained for each of the noise conditions, by the new posterior union model, compared to the conditional union model, the product model, and the baseline full-band HMM. The accuracy rates for the conditional union model and the baseline HMM are quoted from [12]. No noise reduction technique was implemented in the baseline model due to the difficulty caused by the nonstationary nature of the noise.

Table 1 indicates the posterior union model improved upon the conditional union model throughout all test conditions, with more significant improvement in low SNR conditions. These improvements are due to the frame-by-frame order estimation implemented in the posterior union model, which enhances the capability of the model for dealing with nonstationary noise. The conditional union model assumed a constant order for all frames, and its performance was thus compromised by the time-varying noise characteristics. Table 1 also indicates that both union models significantly outperformed the product model and the full-band model, neither of these showing significant robustness to the noise corruption. Figure 2 presents a summary of the results for the four systems, showing the string accuracy as a function of SNR, averaged over all the four noise types.

Improved performance was also obtained for the new model in stationary band-selective noise. The noise was additive, and simulated by passing Gaussian white noise through a band-pass filter. The central frequency and bandwidth of the noise were varied to create the effects that there were one subband, two subband, and three subband corruption, respectively, within the five subbands of the system. A total of eight different noise conditions were generated, including three cases with one subband corruption (affecting subbands 2, 3, and 4, resp.), three cases with two subband corruption (affecting subbands 2 and 3, 3 and 4, and 4 and 5, resp.), and two cases with three subband corruption (affecting subbands 2, 3, and 4, and subbands 3, 4, and 5, resp.). With the above knowledge about the noise, we implemented an "ideal" conditional union model for comparison. The model, based on (10), used a fixed order M over the duration of each test utterance that matched the number of noisy subbands in the utterance. The matched orders were derived from the prior knowledge of the structure of the noise with additional manual refinement to optimize the performance against the order. Table 2 shows the string accuracy, averaged over all the eight noise conditions, obtained by various models. Figure 3 shows the histograms of the orders selected by the posterior union model and the conditional union model in the above noise conditions. The conditional union model selected the orders based on the state-occupancy match, which is a sentence-level statistic involving a balance across all the frames within the sentence. As a result, the conditional union model matched the sentence-level average noise information better than the posterior union model, as indicated by the higher peaked histograms for the conditional union model, at the orders correctly reflecting the numbers of noisy subbands within the test sentences. However, the posterior union model exploited the frame-level SNR more effectively. In stationary noise, the number of useful subbands can still change from frame to frame due to the time-varying speech spectra and hence the time-varying frame/subband SNR. Figure 4 presents an example, showing the order sequence produced by the posterior union model for an utterance with one subband corruption at SNR = 10 dB. For the high SNR frames, the model tended to choose a low order to keep the high SNR subbands in recognition, whilst for the low SNR or noise-dominated frames, the model tended to choose a high order to remove the noise-affected subbands from recognition. The better exploitation of the local SNR for order selection may account for the improved performance for the posterior union model. In our experiments the manually optimized fixed order model remained the best, as shown in Table 2, indicating that there is still room for improvement over the order estimation.

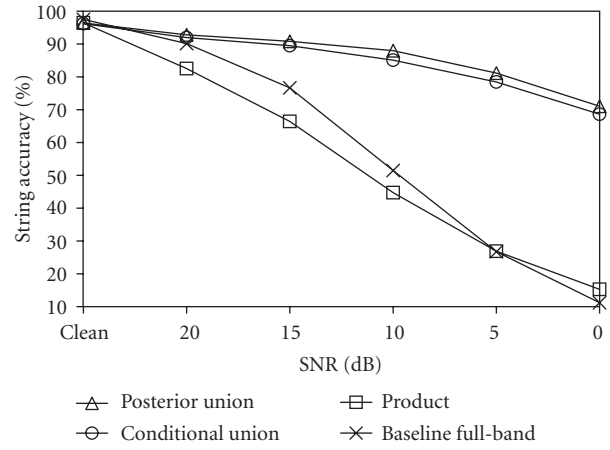


Figure 2: String accuracy as a function of SNR, averaged over four real-world noises (telephone ring, whistle, contact, and connect), for the posterior union model, conditional union model, product model, and baseline full-band HMM.

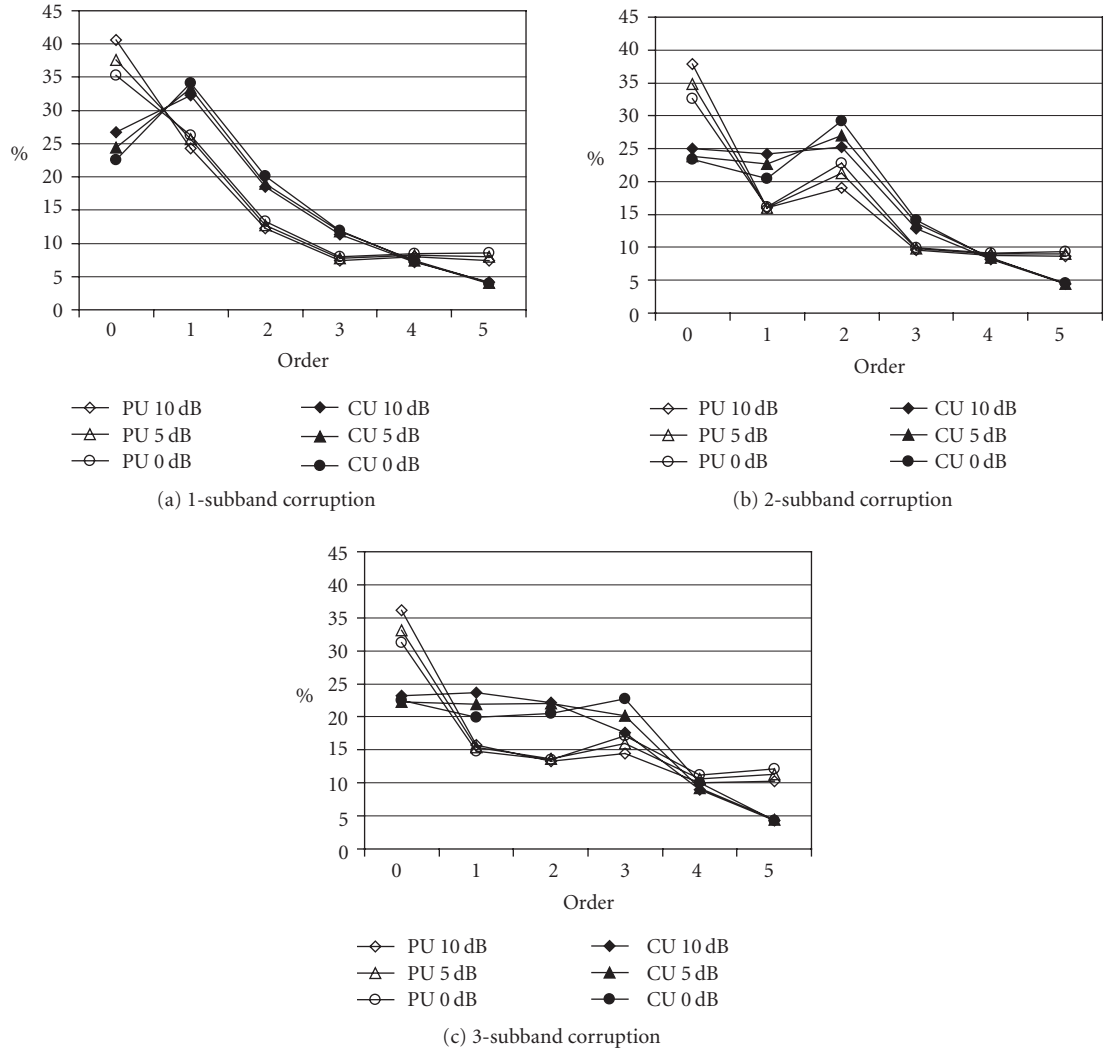
4.2. Experiments on SPIDRE for speaker identification

As shown above, the state-occupancy method, which is based on the statistics of the number of speech frames assigned to each individual HMM state, may be used to estimate the order for a conditional union model, when the model is incorporated into a multistate HMM for applications such as speech recognition. However, this method is invalid for an HMM with the use of only a single state to account for all the frames, for example, a GMM, which has been widely used for speaker recognition. This subsection describes the use of the posterior union model for speaker identification. The new model estimates the order on a frame-by-frame basis and can be applied to a single-state HMM or GMM. The model is defined in (13) and (14), and uses subband features to model speech subject to unknown, time-varying band-selective corruption.

[3] The string accuracy is used to measure the performance, that is, a test utterance is correctly recognized if all digits in the utterance are correctly recognized, without insertion and deletion.

Table 2: Average digit string accuracy (%) in stationary band-selective noise, for the posterior union model, compared with the conditional union model, the product model, the union model with manually optimized order matching the number of noisy bands, and the baseline full-band HMM.

SNR (dB)	Posterior union	Conditional union	Product model	Matched order	Baseline full-band
20	93.87	93.80	83.46	94.23	87.91
15	92.91	92.12	66.48	93.88	74.67
10	92.45	89.90	44.52	92.72	52.99
5	89.33	86.33	27.12	91.49	29.97
0	83.47	80.91	15.75	85.79	13.93



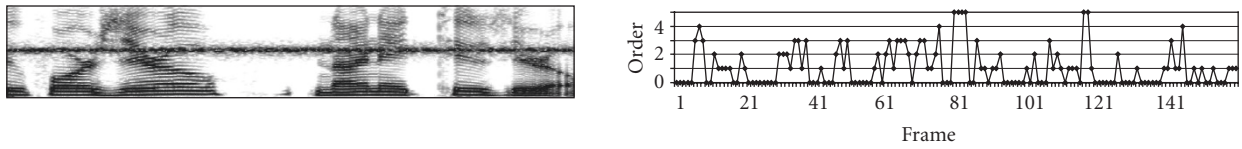
(a) 1-subband corruption

(b) 2-subband corruption

(c) 3-subband corruption

Figure 3: Histograms of the orders selected by the posterior union model (PU) and conditional union model (CU), in stationary band-selective noise with 1-subband, 2-subband, and 3-subband corruption within 5 subbands modeled by 10 feature streams (5 static and 5 delta subband cepstra), at 10 dB, 5 dB, and 0 dB SNRs.

The SPIDRE database [17], a subset of the Switchboard corpus designed for speaker identification research, was used in the experiments. The database contains 45 target speakers (27 male, 18 female). For each speaker, four conversation halves are provided (denoted by A1, A2, B, C), which originate from three different handsets with two conversations (A1, A2) from the same handset. In our experiments, we trained the model for each speaker on two conversations



(a) (b)

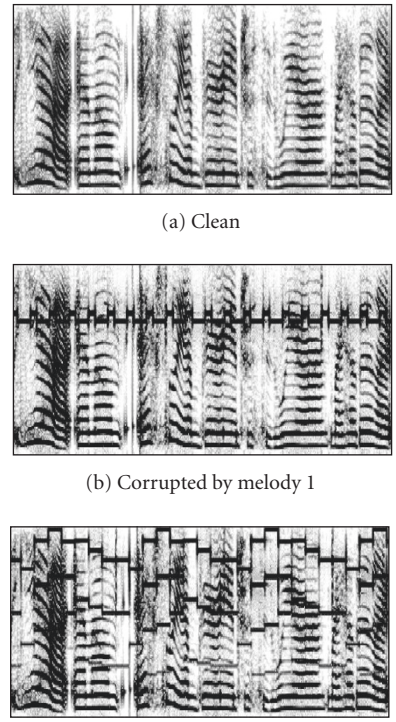
Figure 4: Order sequence (b) produced by the posterior union model, for an utterance with 1-subband corruption at SNR = 10 dB (a).

(A1, B), and tested on one matched conversation (A2, handset used in training data) and one mismatched conversation (C, handset not used in training data). Each conversation half has approximately two minutes of speech. The first 15 seconds of speech from each test conversation was used for test utterances. This experimental setup is similar to that described in [18]. Previous studies on the database were focused on the effects of handset variability. This study is focused on the effect of noise. Earlier research for speaker recognition has targeted the impact of background noise through filtering techniques such as spectral subtraction or Kalman filtering [19, 20]. Other techniques rely on a statistical model of the noise, for example, parallel model combination (PMC) [21, 22]. The missing-feature method has been studied in [3, 6], showing improved robustness by ignoring the strongly distorted feature components. The posterior union model represents an alternative to the missing-feature method, without assuming identify of the corrupted components.

The speech was divided into frames of 20 ms at a frame period of 10 ms. A new type of subband features, different from the subband MFCC as used in Section 4.1, was used in the speaker identification experiments. The new features were obtained by decorrelating the log filter-bank amplitudes using a high-pass filter $H(z) = 1 - z^{-1}$. As suggested in [23, 24], the filtered log filter-bank amplitudes may be used as an alternative to the conventional MFCC for speech recognition. This feature format is particularly flexible in forming the subband features. Specifically, for each frame a 13-channel, band-limited (300–3100 Hz) mel-scale filter bank was used to obtain 13 log filter-bank amplitudes. These were decorrelated using the high-pass filter into 12 decorrelated log filter-bank amplitudes, denoted by $D = (d_1, d_2, \ldots, d_{12})$ (the time index for the frame is omitted for clarity). Vector D can be viewed as a frame vector consisting of 12 independent subband components, and thus be modeled by the union model. The bandwidth of the subband can be conveniently increased by grouping neighboring subband components together to form a new subband component. For example, D can be converted into a 6-subband frame vector by grouping every two consecutive components into a new component, that is,

$$D = (\{d_1, d_2\}, \{d_3, d_4\}, \ldots, \{d_{11}, d_{12}\}) \longrightarrow X = (x_1, x_2, \ldots, x_6), \tag{16}$$

where each x_n contains two decorrelated log amplitudes corresponding to two consecutive filter-bank channels. The new frame vector X contains subband components each covering a wider frequency range than the subband components in D. This 6-subband vector, with the subtraction of the sentence-level mean (similar to cepstral mean removal) and with the addition of the delta vector, was used in the experiments. Thus, there was a feature vector of twelve streams, six static and six dynamic, for each frame. This frame vector was modeled by the posterior union model with $N = 12$ and an order range $0 \leq M \leq 6$, allowing up to six stream corruption. For comparison, a product model and a baseline recognition system based on GMM were implemented. The product model used the same features as the union model and the baseline GMM used a full-band feature vector of the same size (12 MFCC plus 12 delta MFCC) for each frame, with the same band limitation and cepstral mean subtraction. All models used 32 Gaussian mixtures with diagonal covariance matrices for each speaker.



(a) Clean

(b) Corrupted by melody 1

(c) Corrupted by melody 2

Figure 5: Spectra of clean and noisy test utterances used in speaker identification experiments.

Table 3: Speaker identification accuracy (%) using clean and noisy utterances with melody 1 noise, for matched (Mat), mismatched (Mis), and combined (Cmb) handset tests.

SNR	Posterior union			Product model			Baseline GMM		
(dB)	Mat	Mis	Cmb	Mat	Mis	Cmb	Mat	Mis	Cmb
Clean	84.44	77.78	81.11	86.67	73.33	80.00	86.67	73.33	80.00
20	82.22	68.89	75.55	73.33	64.44	68.88	77.78	68.89	73.33
15	80.00	66.67	73.33	66.67	62.22	64.44	73.33	64.44	68.88
10	77.78	66.67	72.22	64.44	46.67	55.55	71.11	62.22	66.66

Table 4: Speaker identification accuracy (%) using noisy utterances with melody 2 noise, for matched (Mat), mismatched (Mis), and combined (Cmb) handset tests.

SNR	Posterior union			Product model			Baseline GMM		
(dB)	Mat	Mis	Cmb	Mat	Mis	Cmb	Mat	Mis	Cmb
20	80.00	75.56	77.78	77.78	57.78	67.78	80.00	64.44	72.22
15	75.56	66.67	71.11	57.78	46.67	52.22	66.67	53.33	60.00
10	66.67	57.78	62.22	44.44	26.67	35.55	48.89	35.56	42.22

Two mobile phone ring noises, labelled as *melody 1* and *melody 2*, were used to corrupt the test utterances. These noises were added, respectively, to each of the test utterances to simulate real-world noise corruption. Both noises exhibit a time-varying nature, especially for melody 2. Figure 5 shows examples of the noisy speech utterances used in the recognition.

Tables 3 and 4 present the identification results in melody 1 and melody 2, respectively, produced by various models as a function of SNR, for the matched, mismatched, and combined handset tests. The posterior union model indicated improved robustness to both noise corruption and handset mismatch in all tested noisy conditions except for one condition, with the melody 1 noise, SNR = 20 dB, and mismatched handset, in which the new model achieved the same accuracy as that by the baseline model. In the clean condition with the matched handset, the new model also experienced a slight loss of accuracy in comparison to the other two models.

5. CONCLUDING REMARKS

This paper described a new statistical method—the posterior union model, for speech and speaker recognition involving partial feature corruption assuming no knowledge about the noise characteristics. The new model is an extension of our previous union model from a conditional-probability formulation to a posterior-probability formulation. The new formulation has potential to outperform the previous conditional union model when dealing with nonstationary noise corruption, as indicated by the experimental results for digits recognition obtained on the TIDIGITS database. The new formulation also offered an approach to incorporate the union model into GMM-based speaker recognition, as demonstrated by the experiments for speaker identification conducted on the SPIDRE database. Compared to the conditional union model, the major part of the additional computation required by the posterior union model is the formation of the posteriors from the likelihoods, which involves the normalization of the likelihoods over all possible candidates for all concerned orders. Our experiments indicate the relative processing time 1/6.3/6.9 for the baseline full-band HMM, conditional union model, and posterior union model for recognizing the 6196 TIDIGITS test utterances.

As with other missing-feature methods, the posterior union model is only effective given partial noise corruption, a condition that cannot be realistically assumed for many real-world problems. Our recent research focused on the extension of the union model for dealing with *full* noise corruption that affects all time-frequency regions of the speech representation. This could be achieved by combining the union model with conventional noise-robust techniques such as noise filtering or multicondition training. Due to lack of knowledge or the time-varying nature of the noise, the conventional techniques for noise removal may only partially clean the speech. The residual noise leftover by an inaccurate noise-reduction processing can be dealt with by the missing-feature methods or by the union model. This may lead to a system that has potential to outperform the individual techniques in isolated operation. Examples of this research, for dealing with broadband noises such as in Aurora 2, can be found in [25, 26].

REFERENCES

[1] R. P. Lippmann and B. A. Carlson, "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise," in *Proceedings of 5th European Conference on Speech Communication and Technology (Eurospeech '97)*, pp. 37–40, Rhodes, Greece, September 1997.

[2] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, vol. 2, pp. 1255–1258, Munich, Germany, April 1997.

[3] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, vol. 1, pp. 121–124, Seattle, Wash, USA, May 1998.

[4] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-band speech recognition in noisy environments," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, vol. 2, pp. 641–644, Seattle, Wash, USA, May 1998.

[5] P. Renevey and A. Drygajlo, "Statistical estimation of unreliable features for robust speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, vol. 3, pp. 1731–1734, Istanbul, Turkey, June 2000.

[6] L. Besacier, J. F. Bonastre, and C. Fredouille, "Localization and selection of speaker-specific information with statistical modeling," *Speech Communication*, vol. 31, no. 2-3, pp. 89–106, 2000.

[7] M. L. Seltzer, B. Raj, and R. M. Stern, "Classifier-based mask estimation for missing feature methods of robust speech recognition," in *Proceedings of International Conference on Spoken Language Processing (ICSLP '00)*, Beijing, China, October 2000.

[8] J. Barker, M. P. Cooke, and P. Green, "Robust ASR based on clean speech models: an evaluation of missing data techniques for connected digit recognition in noise," in *Proceedings of 7th European Conference on Speech Communication and Technology (Eurospeech '01)*, pp. 213–217, Aalborg, Denmark, September 2001.

[9] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Communication*, vol. 34, no. 1-2, pp. 25–40, 2001.

[10] M. P. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2001.

[11] J. P. Barker, M. P. Cooke, and D. P. W. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, no. 1, pp. 5–25, 2005.

[12] J. Ming, P. Jancovic, and F. J. Smith, "Robust speech recognition using probabilistic union models," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 403–414, 2002.

[13] J. Ming and F. J. Smith, "Speech recognition with unknown partial feature corruption—a review of the union model," *Computer Speech and Language*, vol. 17, no. 2-3, pp. 287–305, 2003.

[14] P. Jancovic and J. Ming, "A probabilistic union model with automatic order selection for noisy speech recognition," *Journal of Acoustic Society of America*, vol. 110, no. 3, pp. 1641–1648, 2001.

[15] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.

[16] R. G. Leonard, "A database for speaker-indpendent digit recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '84)*, pp. 42.11.1–42.11.4, San Diego, Calif, USA, March 1984.

[17] J. P. Campbell Jr. and D. A. Reynolds, "Corpora for the evaluation of speaker recognition systems," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '99)*, vol. 2, pp. 2247–2250, Phoenix, Ariz, USA, March 1999.

[18] D. A. Reynolds, "The effects of handset variability on speaker recognition performance: experiment on the Switchboard corpus," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '96)*, pp. 113–116, Atlanta, Ga, USA, May 1996.

[19] J. Ortega-Garcia and L. Gonzalez-Rodriguez, "Overview of speaker enhancement techniques for automatic speaker recognition," in *Proceedings of International Conference on Spoken Language Processing (ICSLP '96)*, pp. 929–932, Philadelphia, Pa, USA, October 1996.

[20] Suhadi, S. Stan, T. Fingscheidt, and C. Beaugeant, "An evaluation of VTS and IMM for speaker verification in noise," in *Proceedings of 8th European Conference on Speech Communication and Technology (Eurospeech '03)*, pp. 1669–1672, Geneva, Switzerland, September 2003.

[21] T. Matsui, T. Kanno, and S. Furui, "Speaker recognition using HMM composition in noisy environments," *Computer Speech and Language*, vol. 10, no. 2, pp. 107–116, 1996.

[22] L. P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 1, pp. 457–460, Salt Lake City, Utah, USA, May 2001.

[23] C. Nadeu, J. Hernando, and M. Gorricho, "On the decorrelation of filter-bank energies in speech recognition," in *Proceedings of 4th European Conference on Speech Communication and Technology (Eurospeech '95)*, pp. 1381–1384, Madrid, Spain, September 1995.

[24] K. K. Paliwal, "Decorrelated and liftered filter-bank energies for robust speech recognition," in *Proceedings of 6th European Conference on Speech Communication and Technology (Eurospeech '99)*, pp. 85–88, Budapest, Hungary, September 1999.

[25] J. Ming and F. J. Smith, "A posterior union model for improved robust speech recognition in nonstationary noise," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 1, pp. 420–423, Hong Kong, April 2003.

[26] J. Ming, "Universal compensation—an approach to noisy speech recognition assuming no knowledge of noise," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, pp. 961–964, Montreal, Canada, May 2004.

Ji Ming is a Reader in computer science at the Queen's University Belfast. He received a B.S. degree from Sichuan University, China, in 1982, an M.Phil. degree from Changsha Institute of Technology, China, in 1985, and a Ph.D. degree from Beijing Institute of Technology, China, in 1988, all in electronic engineering. He was Associate Professor with the Department of Electronic Engineering, Changsha Institute of Technology, from 1990 to 1993. From August 2005 to February 2006, he was a Visiting Scientist at the MIT Computer Science and Artificial Intelligence Laboratory. His research interests include speech and language processing, image processing, signal processing, and pattern recognition.

Jie Lin is a Ph.D. candidate in the University of Electronic Science and Technology of China. He received a B.S. degree in computer science and engineering from the same university in 2003, and has recently completed his M.Phil. thesis on robust speech recognition within the university. His main research interests are in pattern recognition, speech and speaker recognition, and computer science.

F. Jack Smith has been a Professor of computer science at Queen's University Belfast since 1997. He received an M.A. degree in physics and a Ph.D. degree in mathematics in 1960 and 1962, respectively, both from Queen's University Belfast. He was Visiting Professor at the University of Connecticut, Storrs, from 1985 to 1986. His research interests are now mainly in artificial intelligence, particularly speech and language processing. Dr. Smith is a Member of the Royal Irish Academy.