ACADEMIC
PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computer Speech and Language 17 (2003) 287–305

COMPUTER
SPEECH AND
LANGUAGE

www.elsevier.com/locate/csl

Speech recognition with unknown partial feature corruption – a review of the union model

Ji Ming [*], F. Jack Smith

School of Computer Science, The Queen's University of Belfast, Belfast BT7 1NN, UK

Received 15 January 2002; received in revised form 14 October 2002; accepted 30 November 2002

Abstract

This paper provides a summary of our studies on robust speech recognition based on a new statistical approach – the probabilistic union model. We consider speech recognition given that part of the acoustic features may be corrupted by noise. The union model is a method for basing the recognition on the clean part of the features, thereby reducing the effect of the noise on recognition. To this end, the union model is similar to the missing feature method. However, the two methods achieve this end through different routes. The missing feature method usually requires the identity of the noisy data for noise removal, while the union model combines the local features based on the union of random events, to reduce the dependence of the model on information about the noise. We previously investigated the applications of the union model to speech recognition involving unknown partial corruption in frequency band, in time duration, and in feature streams. Additionally, a combination of the union model with conventional noise-reduction techniques was studied, as a means of dealing with a mixture of known or trainable noise and unknown unexpected noise. In this paper, a unified review, in the context of dealing with unknown partial feature corruption, is provided into each of these applications, giving the appropriate theory and implementation algorithms, along with an experimental evaluation.

1. Introduction

This paper studies noisy speech recognition assuming that there is no knowledge about the noise, except that the noise is *localized* in certain areas of the temporal-spectral feature space. We term this a partial corruption of speech. More specifically, partial corruption may include partial

[*] Corresponding author. Tel.: +44-28-90274723; fax: +44-28-90683890.
E-mail address: j.ming@qub.ac.uk (J. Ming).

0885-2308/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0885-2308(03)00003-2

frequency-band corruption, partial duration corruption, partial feature corruption, and their combinations. Partial frequency-band corruption may be caused by frequency-selective noise, for example, a telephone ring, a whistle, a siren or a random channel tone, which usually have a band-limited characteristic and thus affect only certain parts of the speech frequency band. Partial duration corruption may be caused by time-limited (but not necessarily band-limited) noise, for example, shut doors, car horns, random channel impulses or any type of burst noise occurring during the utterance and affecting only certain parts of the temporal signal. In addition, partial corruption may also occur within the feature components or feature streams, when some of the components or streams are more sensitive than the others to a certain type of noise. For example, the static cepstral features are more adversely affected by convolutional noise (i.e., channel effect) than the dynamic (i.e., delta) cepstral features.

There may be two different ways to deal with partial corruption for speech recognition. Firstly, we may use the conventional noise-reduction techniques to remove the noise from the signal, or to adapt the model to the noisy observations. These techniques usually require certain knowledge such as the spectral or cepstral characteristics of the noise, and these can be difficult to estimate if the noise is unpredictable and/or non-stationary. Alternatively, we may base the recognition mainly on information from the clean parts of the features, by ignoring the noisy parts, or by making these parts play a less significant role in recognition. This is the idea of the missing feature method, which indicates that an effective strategy for reducing the effects of noise corruption is to detect and simply ignore strongly affected features (see, for example, Cooke, Morris, & Green, 1997; Lippmann & Carlson, 1997; Drygajlo & El-Maliki, 1998; Vizinho, Green, Cooke, & Josifovski, 1999; de Veth, Cranen, de Wet, & Boves, 1999; Seltzer, Raj, & Stern, 2000; Cooke, Green, Josifovski, & Vizinho, 2001). This method is of interest because no knowledge is required for the noise, except the locations of the affected areas in the feature space. Speech recognition based on partial information is possible due to the redundancy of the temporal-spectral characteristics of speech.

This paper is focused on the second method, i.e., basing the recognition on the subset of features that carry reliable information about the speech utterance. To achieve this, the missing feature method usually requires a labelling of every local 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 on the noise. Recent studies have suggested that the unreliable data may be identified by explicitly measuring the local signal-to-noise ratio (SNR), based on an estimate of the local noise spectrum obtained during a non-speech period (Drygajlo & El-Maliki, 1998; Dupont, 1998; Vizinho et al., 1999). Additionally, Raj, Singh, and Stern (1998) have studied the detection of the corrupted data based on some prior information (e.g., temporal correlation) related to the noise, and Renevey and Drygajlo (2000) have studied this problem based on statistical distribution of the noise, estimated during speech pauses. Based on the identification, further feasibility studies have been carried out towards the reconstruction of the damaged features (e.g., Dupont, 1998; Raj et al., 1998; Josifovski, Cooke, Green, & Vizinho, 1999; Renevey & Drygajlo, 2000). These methods perform well when the corrupting noise is stationary or trainable. For unknown or non-stationary noise, Seltzer et al. (2000) have suggested that some characteristics of the speech signal itself, such as the harmonic nature of voiced speech, may be exploited for identifying the corrupted time-frequency regions; Baker, Cooke, and Green (2001) have studied the combination of auditory scene analysis (e.g., harmonicity) and SNR to produce missing data masks for mixed voiced and unvoiced speech.

In this paper, we introduce a new approach, i.e., the probabilistic union model, for dealing with unknown, time-varying partial corruption. Unlike the missing feature method, the union model does not require the identity of the noisy features; instead, it combines the local features based on the probability theory for the union of random events, to reduce the dependence of the model on information about the noise. We have studied several applications for the union model, including the combination of sub-band information for dealing with partial frequency-band corruption, the combination of local temporal information for dealing with partial duration corruption, and the selection of feature streams for dealing with partial feature stream corruption. In addition, we have studied the combination of the union model with conventional noise-reduction techniques for dealing with a mixture of known or trainable noise and unknown unexpected noise. In this paper, a united review is provided into each of these applications, in the context of speech recognition with unknown partial feature corruption.

This paper is organized as follows. In Section 2 we present the theory of the union model. From Section 3 to Section 6 we introduce its applications, giving the appropriate algorithms and an experimental evaluation. Finally, a summary is given in Section 7, along with some possible future improvements.

2. The union model

Assume that a speech utterance may be represented by a set of N feature vectors $X = (x_1, x_2, \ldots, x_N)$, where each x_n may correspond to a specific feature stream, for a specific time-frequency region. We consider speech recognition based on the conditional probability $P(X|w)$ of X, associated with each candidate speech unit w, where w may be a word or a subword unit. The problem of speech recognition with unknown partial corruption may be expressed as the problem of how to calculate the probability $P(X|w)$, given that some of the feature vectors x_n in X may be noisy, but without knowledge about their identity. When there is no noise, we normally define $P(X|w)$ as the joint conditional probability of all the feature vectors. This is equivalent to combining the x_n using the "and" (i.e., conjunction) operator $\wedge$, assuming that they are all reliable (i.e., present). Assuming independence between the feature vectors, this joint probability equals the product of the individual feature probabilities, i.e.

$$P(X|w) = P(x_1 \wedge x_2 \wedge \cdots \wedge x_N|w) = P(x_1|w)P(x_2|w)\cdots P(x_N|w), \tag{1}$$

where $P(x_n|w)$ is the conditional probability of feature vector x_n given the speech unit w. Eq. (1) has been employed in the hidden Markov models (HMMs), for example, for computing the output probability for multiple feature streams (e.g., static and dynamic feature streams) associated with each state, and for computing the probability for a complete observation (i.e., frame) sequence. The probability $P(X|w)$ produced by Eq. (1) is typically dominated by small probabilities of features, as a result of the product. This characteristic makes the model effective in discriminating between correct and incorrect speech units based on local feature differences, but also makes it sensitive to local feature corruption. When the model is trained on clean speech and is applied to a test utterance with some noisy features, the model probabilities $P(x_n|w)$ for the noisy x_n may become small for the correct speech unit due to the noise caused mismatch between the model and data. When these small and random feature probabilities become

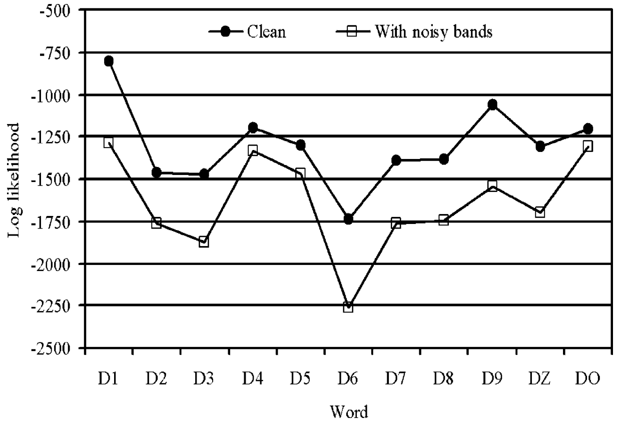


Fig. 1. Likelihood associated with each of the eleven digits D1–DO when D1 is spoken, based on a model in Eq. (1), showing that a partial frequency-band corruption reduces the likelihood of the correct word (SNR = 0 dB).

dominant, the model's trained ability to produce a high probability for the correct speech unit will be destroyed.

As an example, Fig. 1 shows the effect of a partial feature corruption on the model probability, Eq. (1), associated with the correct speech unit. The model shown in the figure uses five independent sub-band feature streams to characterize an utterance, i.e., $X = (x_1, x_2, \ldots, x_5)$, where x_n is the feature stream from the nth sub-band [1]. The speech units being modelled are eleven digits: "one" to "nine", "zero" and "oh". Fig. 1 shows the likelihood associated with each of the 11 words when "one" is spoken (averaged over 226 utterances), assuming no noise and a band-limited noise which corrupts two sub-bands within the five sub-bands, respectively. As indicated in Fig. 1, the partial band corruption leads to a lower likelihood for the correct word "one", and most importantly, to lower likelihood comparisons between the correct word and the competitive words (e.g., "oh" and "four"). These low likelihood comparisons lead to a low recognition rate.

The above problem can be improved by removing the probabilities corresponding to the unreliable features from the product, leaving a marginal conditional probability including only the reliable features in X for recognition. This is the idea of the missing feature method (Cooke et al., 1997; Lippmann & Carlson, 1997), which assumes that the corrupted (i.e., missing) features are known or can be identified. A study by de Veth et al. (1999) has attempted to release this assumption by using a back-off model, in which each feature probability distribution is formed as a weighted combination of two distributions: one from the training data and another being a uniform distribution, to account for possible outliers arising from an unknown noise. The union model described below represents an alternative method for getting around the corrupted features assuming no knowledge about their identity.

The union model deals with the uncertainty of the corrupted features by using the "or" (i.e., disjunction) operator to combine the subsets of the features, to assume that *any* of the subsets may be the set of reliable features. The probability based on a union model for a feature set X with N feature vectors may be written in a general form as

[1] The model will be discussed in full detail in Section 3.

$$P(X|w) = P\left(\bigvee_{n_1 n_2 \cdots n_{N-M}} x_{n_1} x_{n_2} \cdots x_{n_{N-M}} | w\right), \tag{2}$$

where the symbol $\vee$ represents the “or” operator, $x_{n_1} x_{n_2} \cdots x_{n_{N-M}}$ is a subset in X containing $(N - M)$ feature vectors which are combined with the “and” operator (for simplicity, we have omitted the symbol $\wedge$ between the x_n), and the “or” operator $\vee$ is applied between all possible subsets of $(N - M)$ feature vectors in X, giving a total of ${}^N C_{N-M}$ combinations. The parameter M in Eq. (2) is called the order of the model, which takes a value in the range $0 \leqslant M \leqslant N - 1$. For example, in the case with four feature vectors (x_1, x_2, x_3, x_4), the union model probability may take four possible forms, corresponding to $M = 0, 1, 2$ and 3, respectively:

$$(M = 0)\ P(X|w) = P(x_1 x_2 x_3 x_4 | w),$$

$$(M = 1)\ P(X|w) = P(x_1 x_2 x_3 \text{ or } x_1 x_2 x_4 \text{ or } x_1 x_3 x_4 \text{ or } x_2 x_3 x_4 | w),$$

$$(M = 2)\ P(X|w) = P(x_1 x_2 \text{ or } x_1 x_3 \text{ or } x_1 x_4 \text{ or } x_2 x_3 \text{ or } x_2 x_4 \text{ or } x_3 x_4 | w)$$

$$(M = 3)\ P(X|w) = P(x_1 \text{ or } x_2 \text{ or } x_3 \text{ or } x_4 | w).$$

Form $(M = 0)$ corresponds to Eq. (1), suited to the situation in which all features are present (i.e., no corruption). Form $(M = 1)$ is suited to the situation in which we know there is one feature being corrupted (i.e., missing) but do not know which, so the reliable features may be any of the combinations of three features. In a similar way, Forms $(M = 2)$ and $(M = 3)$ are suited, respectively, to the situations in which we know there are two and three features being corrupted, but have no information about the identities of the missing features.

The expression for the probability of the union model is readily derived using the rules of probability for the union of random events [2]. To illustrate how the probability may be calculated and how the model may get around the corrupted features, use the above example for four feature vectors with order $M = 2$, assuming two corrupted feature vectors with an unknown identity. Assuming independence between the feature vectors, the probability $P(X|w)$ in Form $(M = 2)$ can be written, based on the rules described in Footnote 2, as

$$\begin{aligned} P(X|w) \simeq\ & P(x_1|w)P(x_2|w) + P(x_1|w)P(x_3|w) + P(x_1|w)P(x_4|w) + P(x_2|w)P(x_3|w) \\ & + P(x_2|w)P(x_4|w) + P(x_3|w)P(x_4|w), \end{aligned} \tag{3}$$

where we have omitted the terms corresponding to the joint probabilities between the combinations $x_i x_j$ for simplicity [3]. Eq. (3) is effectively the sum of the products of the probabilities of all

[2] The probability of the union of two random events $P(A \text{ or } B) = P(A) + P(B) - P(AB)$, or $P(A) + P(B) - P(A)P(B)$ if A and B are independent, and $P(A) + P(B)$ if they are exclusive. The probability of the union of multiple events $A_1, A_2, \ldots, A_N$ may be obtained by using a recursion $P(\vee_{i=1}^{n} A_i) = P(\vee_{i=1}^{n-1} A_i) + P(A_n) - P((\vee_{i=1}^{n-1} A_i) \wedge A_n)$, which is also equal to $P(\vee_{i=1}^{n-1} A_i) + P(A_n) - P(\vee_{i=1}^{n-1} A_i)P(A_n)$ if the A_n are mutually independent, $n = 2, \ldots, N$.

[3] We have ignored the terms corresponding to $P(\vee_{i=1}^{n-1} A_i)P(A_n)$ in $P(\vee_{i=1}^{n-1} A_i) + P(A_n) - P(\vee_{i=1}^{n-1} A_i)P(A_n)$, which are the joint probabilities between the events of union. These probabilities are typically small in comparison to the probabilities of the individual events, and are equal to zero when the events are exclusive. In the above example, when the number of corrupted features equals 2, the subsets combining two different features are exclusive to one another in terms of containing the remaining two reliable features.

possible combinations between two feature vectors, which therefore includes the product of the probabilities of the remaining two "clean" feature vectors, providing correct information about the speech unit. A major difference between Eqs. (3) and (1) is that Eq. (3) is not dominated by small $P(x_i|w)P(x_j|w)$, but by large $P(x_i|w)P(x_j|w)$. If we can assume that the two clean feature vectors produce a large product (i.e., joint probability) $P(x_i|w)P(x_j|w)$ for the correct speech unit, then selecting the maximum union probability $P(X|w)$ for recognition has a chance to get the correct speech unit without requiring the identity of the two corrupted feature vectors. We may further assume that the discriminative power of the model is proportional to the size of the product (i.e., the number of factors in the product) corresponding to the clean features.

In general, if the number of noisy feature vectors is M, then the union model with order M (i.e., Eq. (2)) is optimal in terms of maximizing the size of the joint probability containing clean feature vectors (which has a number of $N - M$). This model may also be used for the situations in which the number of corrupted features is less than M (for example, Eq. (3) may also be used for one noisy feature vector or none, in which cases three or all of the combinations $x_i x_j$ in the union may correspond to clean features). But this may be less effective in terms of discrimination than an order-matched model, because of the marginalization of the joint probability of the clean features. In practice, we need a high order to accommodate as many noisy features as possible, but a low order to retain discrimination for clean speech recognition, i.e., seeking a balance between noise robustness and clean speech performance. This principle has been studied in our experiments as a possible method for selecting the order for the union model, to offer robustness against uncertainty on the number of corrupted features. In addition, an algorithm for automatic selection of the model order has been studied for the combination of sub-band features, discussed next.

3. Application to sub-band recombination

3.1. Model and algorithms

Sub-band based speech recognition has been studied by Bourlard and Dupont (1996) and by Hermansky, Tibrewala, and Pavel (1996), as an alternative to the traditional full-band based method, for dealing with band-limited noise corruption. In the sub-band approach, the full speech frequency band is divided into several sub-bands, and each sub-band is featured independently of the other sub-bands, so that a local frequency-band corruption will not spread over the entire feature space. For recognition, a critical issue is how to formulate the recombination of the sub-band features, given no prior knowledge about the noisy sub-bands. Several methods have been studied for sub-band recombination, typically, including linear combination, neural combination, full combination, and graphical models (see, for example, Bourlard & Dupont, 1996; Hermansky et al., 1996; Cerisara, Haton, Mari, & Fohr, 1998; Okawa, Eorico, & Potamianos, 1998; Mirghafori & Morgan, 1998; Morris, Hagen, & Bourlard, 1999; Daoudi, Fohr, & Antoine, 2000).

The union model may be used for sub-band recombination (Ming & Smith, 1999, 2000, 2001). Now the feature set $X = (x_1, x_2, \ldots, x_n)$ is a collection of feature vectors from N sub-bands, with x_n being the feature vector from the nth sub-band. Assume that some of the x_n may be corrupted due to an unknown band-limited noise. Consider using the union model to estimate the state emission probability in an HMM. Denote by $X(t) = (x_1(t), x_2(t), \ldots, x_N(t))$ the sub-band feature set at frame time t and by $X_1^T = (X(1), X(2), \ldots, X(T))$ an observation sequence of T frames. An HMM for the probability of X_1^T may be written as

$$P(X_1^T|\lambda) = \sum_S P(S|\lambda) \prod_{t=1}^{T} B_{s_t}(X(t)), \tag{4}$$

where $P(S|\lambda)$ is the probability of the state sequence S, and $B_i(X)$ is the emission probability for frame feature set X in state i based on a union model, i.e.

$$B_i(X) = B_i\left(\bigvee_{n_1 n_2 \cdots n_{N-M}} x_{n_1} x_{n_2} \cdots x_{n_{N-M}} \right). \tag{5}$$

As described earlier in Footnote 3, we may assume that in the union probability the terms corresponding to the joint probabilities between the events of union are small and can be neglected in comparison to the probabilities of the events themselves. Therefore, assuming independence between the sub-band feature vectors, $B_i(X)$ can be approximated as

$$\begin{aligned} B_i(X) &\simeq \sum_{n_1 n_2 \cdots n_{N-M}} B_i(x_{n_1}) B_i(x_{n_2}) \cdots B_i(x_{n_{N-M}}) \\ &\propto \sum_{n_1 n_2 \cdots n_{N-M}} b_i(x_{n_1}) b_i(x_{n_2}) \cdots b_i(x_{n_{N-M}}), \end{aligned} \tag{6}$$

where $B_i(x_n)$ is the emission probability distribution for vector n in state i, $b_i(x_n)$ is the corresponding emission probability density, and the summation is over all possible combinations of $(N - M)$ feature vectors taken from X. Eq. (6) indicates that we may use the likelihoods, instead of probabilities, to approximate a union probability in a continuous-observation HMM.

The model defined in Eq. (4) can be trained efficiently based on clean data. Although $B_i(X)$ varies with the order M for recognition, there is only one form, with order $M = 0$, that best matches a clean feature set with all features being present. Therefore in the training stage we can compute the union emission probability $B_i(X)$ as the full conjunction probability $B_i(x_1) \cdots B_i(x_N)$, which is proportional to the likelihood $b_i(x_1) \cdots b_i(x_N)$. So the training involves the maximization of the probability $P(X_1^T|\lambda)$ with $B_i(X)$ replaced by $b_i(x_1) \cdots b_i(x_N)$. This maximization can be accomplished efficiently by using the standard forward–backward re-estimation algorithm.

In recognition, we need to decide the order M for computing the union emission probability based on Eq. (6). As described in Section 2, given no knowledge about the number of corrupted sub-bands, we may select a high order to accommodate as many noisy bands as possible, subject to an acceptable performance for clean speech recognition. We call this the balanced fixed-order algorithm. However, an algorithm for automatic order selection is available for the problem of sub-band combination, by exploiting the difference of the state durational structure between the models with and without a matched order.

As discussed in Section 2, a model with a matched order (i.e., an order equaling the number of corrupted features) maximizes the size of the joint probability containing the clean features. As such, we can assume that this model exhibits more characteristics of a clean model versus clean speech, than the model with a mismatched order does. In particular, we assume that an order-matched union model produces a state sequence that is closer to a clean-utterance state sequence, than the state sequences produced by the order-mismatched models. Similar state sequences have similar state durational structures. Therefore, an estimate of the matched order can be obtained by selecting, from a range of orders, the order that produces a state duration structure that is most similar to the state duration structure obtained for the clean utterances. Specifically, denote by $P_i^w(d)$ the state duration probability, for d frames in state i of speech unit w, which is estimated in the training stage using the clean training data. Given a test utterance, we perform recognition with a range of orders M (typically, $0 \leqslant M < N - 1$, where N is the number of the sub-bands), assuming that these will include the matched order. For each order, we obtain a recognition result (in the form of a unit sequence) $W(M) = w_1(M)w_2(M) \cdots w_{k_M}(M)$, along with the associated state duration $d_i(M)$, for each state i of $W(M)$. An estimate of the matched order is given by the order whose associated state duration has the maximum probability, i.e.

$$\hat{M} = \arg\max_M \frac{1}{S(M)} \sum_{w \in W(M)} \sum_{i \in w} \ln P_i^w(d_i(M)), \tag{7}$$

where $S(M)$ stands for the total number of states in $W(M)$. The final recognition result is then given by $W(\hat{M})$. This algorithm is an extension of an earlier algorithm (Jancovic & Ming, 2001a) from isolated-word recognition to connected-word/continuous speech recognition.

3.2. Experimental evaluation

The above sub-band union model has been applied to speech recognition involving unknown partial frequency-band corruption. In the following we summarize the experimental results based on the TIDigits database (Leonard, 1984) for speaker-independent connected digit recognition. The database contained utterances from 225 adult speakers, divided into training and testing sets. The test set provided 6196 utterances from 113 speakers for connected digit recognition. 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.

The speech was sampled at 8 kHz, and was divided into frames of 256 samples. For each frame, we used a mel-scaled filter bank to obtain the log-amplitude spectra (i.e., log FB spectra) of speech. These log FB spectra were then uniformly grouped into sub-bands. For each sub-band, we calculated the sub-band mel-frequency cepstral coefficients (MFCCs) plus the first-order delta MFCCs as the feature vector. In particular, we built 5 sub-bands from a 30-channel filter bank, with each sub-band being modelled by using three MFCCs plus three delta MFCCs. Thus, for this 5-band system, the overall size of the feature set for a frame is $5 \times 6 = 30$. For comparison, we also implemented a baseline HMM which used a full-band feature vector, consisting of 10 MFCCs and 10 delta MFCCs, for each frame. In the experiments, each digit was modelled by a left-to-right HMM with 10 states without state skipping, and each state consisted of 8 Gaussian mixture densities with diagonal covariance matrices. Both the union model and the baseline HMM were trained on clean training data. In training, we recorded the histograms of state occupancy for each digit, as the estimates of the state duration probabilities. The state duration probability was used by the union model for selecting the model order in test, as described in Section 3.1.

We first tested the models with corruption by simulated band-selective noise. The noise was additive, and was generated 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 sub-band, two sub-band and three sub-band corruption, respectively, within the five sub-bands of the system. A total of eight different noise conditions were generated, including three cases with one sub-band corruption (affecting sub-band 2, 3 and 4, respectively), three cases with two sub-band corruption (affecting sub-bands 2 and 3, 3 and 4, and 4 and 5, respectively), and two cases with three sub-band corruption (affecting sub-bands 2, 3, and 4, and sub-bands 3, 4, and 5, respectively). The SNR was calculated for each utterance as the comparison between the average power of speech and the average power of noise within the utterance.

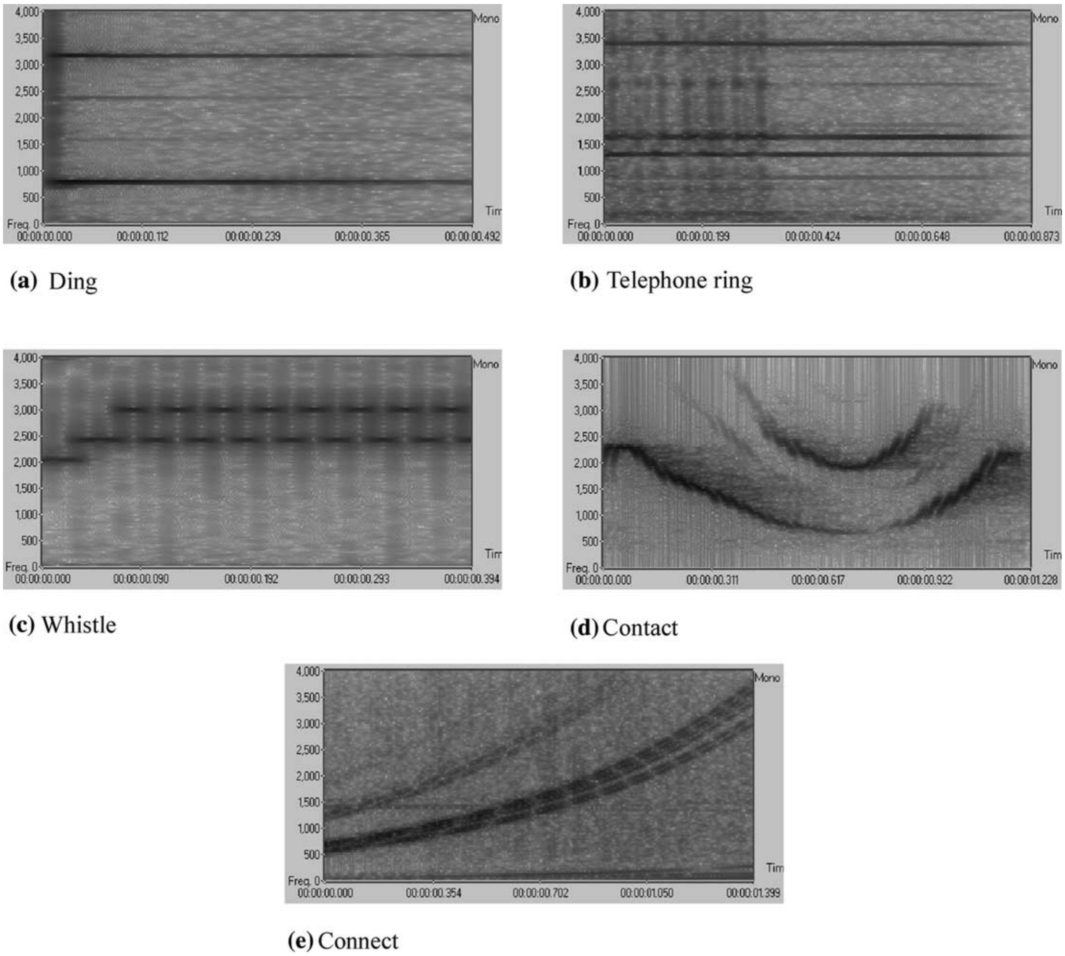
Table 1 presents the string accuracy obtained by various models as a function of the SNR, averaged over the above noise conditions. In Table 1, three models, i.e., the union model with automatic order selection, the model in Eq. (1) (equivalent to the union model with order 0) and the baseline full-band HMM, assumed no knowledge about the noise. These are compared with three other models with knowledge of the noise: (1) the “oracle” model, an ideal missing-feature model, which assumed full a priori knowledge of the corrupted sub-bands and removed those bands manually from the recognition; (2) the union model with an order matching the number of corrupted sub-bands; and (3) a full-band HMM equipped with a Wiener filtering front-end for removing the noise, based on an estimate of the noise spectrum in the interval without speech.

Table 1 shows that the model in Eq. (1) lacked robustness to local band corruption. As expected, the oracle model performed better than the union model. However, in the cases with a high SNR (e.g., 10 dB), the union model achieved a better average performance. This is because throwing away the sub-bands with relatively high SNR in the oracle model caused a loss of useful information. Table 1 also indicates that the accuracy based on the automatically selected order was close to the accuracy with a matched order. Besides, Table 1 indicates that the Wiener filtering considerably improved the performance of the baseline model.

Further tests were conducted for the union model, with automatic order selection, for corruption by real-world noise. The noise data used in the experiments are shown in Fig. 2, which include the sounds of a ding, a telephone ring, a whistle, and the sounds of “contact” and “connect”, extracted from an Internet application. These noises each have a dominant band-selective characteristic, and the noises “contact” and “connect” are particularly non-stationary. These noises were added, respectively, to each of the test utterances and lasted from the beginning to the end of the utterance (the noise was repeated if it shorter than the speech utterance). Table 2 presents the average string accuracy obtained for these noises, by the union model and by the full-band HMM. No noise-reduction technique was implemented in the full-band HMM due to the difficulty caused by the non-stationary nature of the noise. In the experiments, we found less

Table 1
String accuracy (%) in simulated band-selective noise, for the sub-band union model with automatically selected order (auto) and matched order (matched), for the model in Eq. (1), for the model with full a priori knowledge of the noise (oracle), for the full-band HMM, and for the full-band HMM with Wiener filtering (WF)

SNR (dB)	Union auto	Union matched	Eq. (1)	Oracle	Full-band	Full-band WF
Clean	95.58	96.48	96.48	96.48	97.53	
10	84.52	84.59	44.52	83.56	52.99	75.28
5	79.00	80.72	27.12	82.26	29.97	53.19
0	72.59	75.28	15.75	79.71	13.93	26.83



(a) Ding

(b) Telephone ring

(c) Whistle

(d) Contact

(e) Connect

Fig. 2. Spectrograms of real-world noise data used in sub-band based speech recognition experiments.

Table 2
String accuracy (%) in real-world noise (ding, telephone ring, whistle, contact, and connect), for the sub-band union model with automatic order, and for the full-band HMM

SNR (dB)	Union auto	Full-band
10	81.66	54.23
5	74.13	30.10
0	64.87	13.41

significant performance for the union model in dealing with the telephone-ring noise and "connect" noise, both having a multi-band or wide-band characteristic, in comparison to the other three types of noise. Wide-band noise affects all sub-bands, which therefore violates the noise-localization assumption made in the sub-band method. To deal with both narrow-band noise and wide-band noise, a combination of different techniques may be needed. Some examples will be discussed later.

4. Application to feature-stream selection

In speech recognition, a speech utterance may be represented by multiple feature streams, each providing complementary information about the utterance. Typical examples are the static and dynamic spectral features, captured over varying time scales (e.g., Dupont & Bourlard, 1997; Wu, Kingsbury, & Morgan, 1998). Multi-resolution features have also been studied in the frequency domain (McCourt, Vaseghi, & Narte, 1998; Hariharan, Kiss, Viikki, & Tian, 2000). Multi-stream may also consist of heterogeneous features produced by different signal processing methods (Christensen, Lindberg, & Andersen, 2000). In real-world applications, due to the background noise or channel effects, there may be only a subset of the given feature streams that remain reliable at a time. For example, the static spectral features are usually more sensitive to channel distortion or slowly varying noise than the dynamic spectral features. If a feature stream is adversely affected, it should play a less significant role than the other unaffected streams in recognition. However, without prior knowledge of the environmental or noise condition, it can be difficult to decide which subset of the feature streams provides reliable information. This uncertainty may be dealt with by using the union model (Ming, Jancovic, Hanna, Stewart, & Smith, 2000).

As an example, we have generalized the sub-band union model, described in Section 3, by applying the union not only to the sub-bands, but also to the the static and dynamic feature streams, to select the feature stream within each sub-band that is least affected by noise. Specifically, we separated the static feature and dynamic feature within each sub-band into two feature streams x_n and Δx_n, where Δx_n represents the dynamic feature stream (i.e., ΔMFCCs). So the frame at time t was represented by a feature set $X(t) = (x_1(t), \ldots, x_N(t), \Delta x_1(t), \ldots, \Delta x_N(t))$ with $2N$ vectors. Then we modelled this frame sequence using the union model in Eq. (4), with an order range $0 \leqslant M \leqslant 2N - 1$. Based on the previously defined 5-band system, we obtained a union model with 10 input feature streams (five for MFCCs and five for ΔMFCCs, each consisting of three components for each frame) and a full order range $0 \leqslant M \leqslant 9$. Using this generalized union model, we repeated all the experiments in Tables 1 and 2. The results obtained by the generalized union model with automatic order selection are shown in Tables 3 and 4. For comparison, the two tables also include the results obtained by the previous union model as shown in Tables 1 and 2. Both Tables 3 and 4 show positive performance improvement for the generalized union model. Positive

Table 3
Improved string accuracy and error reduction (%) in simulated band-selective noise, by the generalized sub-band union model with feature-stream selection, in comparison to the previous sub-band union model in Table 1, both with automatic orders

SNR (dB)	Union model		Error reduction
	Previous	Generalized	
Clean	95.58 (5922)	96.21 (5961)	14.3 (39)
10	84.52 (5237)	89.90 (5570)	34.8 (333)
5	79.00 (4895)	86.33 (5349)	34.9 (454)
0	72.59 (4498)	80.91 (5013)	30.4 (515)

The absolute numbers of correctly recognized utterances and error reduction are included in brackets, for 6196 test utterances.

Table 4
Improved string accuracy and error reduction (%) in real-world noise, by the generalized sub-band union model with feature-stream selection, in comparison to the previous sub-band union model in Table 2, both with automatic orders

SNR (dB)	Union model		Error reduction
	Previous	Generalized	
10	81.66 (5060)	86.25 (5344)	25.0 (284)
5	74.13 (4593)	80.37 (4980)	24.1 (387)
0	64.87 (4019)	71.85 (4452)	19.9 (433)

The absolute numbers of correctly recognized utterances and error reduction are included in brackets, for 6196 test utterances.

performance gain for the generalized model has also been observed in a recent informal test, in which the model trained on the TIDigits database was used for some local speakers with both channel and accent differences. The improvements may be due to the separation and deemphasis of those static features that were severely affected by the noise or by the channel. A further example for combining feature streams based on the union model has been discussed in a recent conference paper (Jancovic & Ming, 2001b), in which two sub-band feature streams of different kinds, one being the MFCCs and another being the frequency-filtering coefficients (FFCs) (Nadeu, Hernando, & Gorricho, 1995), were combined within the model in Eq. (4). The FFCs were found to be more robust to wide-band noise than the MFCCs. The combined system provided improved robustness for wide-band noise corruption within the sub-band framework.

In the above experiments we have assumed that each frame was modelled by static MFCCs plus delta MFCCs. This somewhat simplified front-end was used to demonstrate the principles of the algorithms. We have now applied the algorithms to a more standard front-end including delta–delta features. We have repeated some of the experiments and obtained improved accuracy. A full description of the new results will be reported later.

5. Application to partial duration corruption

5.1. Model and algorithms

The sub-band method is focused on the noise localized in the frequency band. This study can be extended to the noise localized in the time duration, whereby causing a partial temporal (or partial duration) corruption within a speech utterance. For wide-band corruption, the sub-band method is not an effective strategy. Therefore a study on the handling of temporal breakup (i.e., loss of all frequency information at certain instants during the utterance) may be of interest. In the following we describe a model based on the union principle.

Denote by $O_1^T = (o_1, o_2, \ldots, o_T)$ the temporal observation (i.e., frame) sequence of a speech utterance, where each frame o_t characterizes the temporal spectrum of speech at time t. The presence of a partial duration corruption means that some of the frames o_t are noisy. We then face the problem of how to extract the reliable frames from an observation sequence, assuming no knowledge about the identity of the corrupted frames. We have studied the use of the union model, Eq. (2), for dealing with this problem. To keep the model computationally effective, we first divide each observation sequence $O_1^T = (o_1, o_2, \ldots, o_T)$ into N consecutive segments $X = (x_1, x_2, \ldots, x_N)$, where each segment x_n consists of the same number of consecutive frames, and then apply the union model to the segment sequence X. This can be implemented within an HMM framework. Specifically, the probability of O_1^T based on the union model may be written as

$$P(O_1^T|\lambda) = \sum_S P(S|\lambda)P(X|S,\lambda), \tag{8}$$

where $S = (s_1, s_2, \ldots, s_T)$ is the state sequence associated with the frame sequence O_1^T, $P(S|\lambda)$ is the probability of the state sequence, and $P(X|S,\lambda)$ is the union probability of the segment sequence X given the state sequence S. Assuming independence between the frames (and hence the segments), $P(X|S,\lambda)$ can be expressed as

$$\begin{aligned} P(X|S,\lambda) &= P\left(\bigvee_{n_1 n_2 \cdots n_{N-M}} x_{n_1} x_{n_2} \cdots x_{n_{N-M}} |S,\lambda\right) \\ &\simeq \sum_{n_1 n_2 \cdots n_{N-M}} P(x_{n_1}|S,\lambda)P(x_{n_2}|S,\lambda) \cdots P(x_{n_{N-M}}|S,\lambda) \\ &\propto \sum_{n_1 n_2 \cdots n_{N-M}} p(x_{n_1}|S,\lambda)p(x_{n_2}|S,\lambda) \cdots p(x_{n_{N-M}}|S,\lambda), \end{aligned} \tag{9}$$

where the approximation is based on the same assumption as for Eq. (6). In Eq. (9), $P(x_n|S,\lambda)$ is the probability of the segment x_n given the state sequence, $p(x_n|S,\lambda)$ is the corresponding likelihood, assuming a continuous-observation HMM, and the summation is over all possible combinations of $(N-M)$ segments taken from X. The likelihood of a segment, $p(x_n|S,\lambda)$, can be obtained from the likelihoods of the individual frames, i.e.

$$p(x_n|S,\lambda) = \prod_{o_t \in x_n} b_{s_t}(o_t), \tag{10}$$

where $b_i(o)$ is the frame-based emission probability density in state i. Because local frame corruption within a segment affects the likelihood of the segment (i.e., Eq. (10)), a segment is considered to be noisy if part or all of its frames are noisy.

Eq. (8) is reduced to the standard HMM when the order of the union $M = 0$. Therefore the model can be trained using the standard algorithms on clean data with order $M = 0$ for no frame corruption. The above model has been previously applied to isolated-word recognition (Ming, Stewart, Hanna, & Smith, 1999). In the early experiments for isolated-word recognition, we assumed that the word-based state sequence, required for calculating the union probability based on Eq. (9), can be derived by using the standard Viterbi algorithm, even though there may be some noisy frames in the observation sequence. This method is simple but may not always find the optimal states for the clean frames, as the noise may introduce a distortion in the frame-state alignment. This distortion can become critical in connected-word or continuous speech recognition, because a wrong frame-state alignment may lead to a wrong model (i.e., word or subword) sequence. As a solution, we later suggested a two-pass, n-best rescoring approach (Ming, 2001). In the first pass, the normal HMMs are applied to generate n-best state sequences by using the Viterbi algorithm, assuming that these will include the optimal state sequence for the clean frames. In the second pass, the union model (i.e., Eq. (9)) is applied to the segment probabilities, associated with each hypothesized state sequence, to produce a union probability on which the final recognition decision is based. In rescoring, the ability of the union model for ignoring strongly mismatching data is exploited to reduce the effect of the corrupted segments on recognition. Since it is not straightforward to extend the automatic order selection algorithm, discussed in Section 3 for the sub-band union model, to the above segment union model, we use the balanced fixed-order algorithm, also discussed in Section 3, to decide an order for the model for recognition experiments.

5.2. Experimental evaluation

The above segment union model has been tested for connected digit recognition, also based on the TIDigits database. In the experiments, we assumed that the test utterances each had involved some partial duration corruption, but without knowledge about the times of occurrence and the statistical nature of the noise. To focus on the ability of the model to accommodate temporal corruption, we used full-band features that were subjected to corruption by both narrow-band noise and wide-band noise. The same full-band feature vector and the same HMM topology as used in Section 3.2 were adopted in the experiments.

As shown in Eq. (9), there are two parameters to be decided, i.e., the number of segments for each utterance, N, and the order of the model, M. Because local frame corruption can affect the joint probability of a whole segment, a large N, i.e., dividing each utterance into small segments, should be desirable for confining the noise and leaving more segments to be noise free. With a large N, the computational cost will be higher for computing the ${}^{N}C_{N-M}$ combinations of segments as indicated in Eq. (9). We have tested the model by choosing different lengths for a segment, to search for a balance between the noise resolution and computational efficiency. We found that a segment length around 10 frames (about 160 ms) was suitable for digit recognition. Given the length of the segment, the number of segments N can be variable across utterances with different duration. As such, it is more convenient to calculate the relative order M/N. A relative order of 0.2, for example, may accommodate up to 20% of the segments in each utterance to be corrupted. With the balanced fixed-order algorithm, we select a high order to accommodate as many noisy segments as possible, subject to an acceptable performance for clean speech recognition. For connected digit recognition, we have found in our experiments that a relative order 0.2 provides a good balance. Therefore in the following we use this order for the experiments involving noise corruption. As described earlier, we used an n-best rescoring strategy for continuous speech recognition. In all the experiments, we limited the number of the rescored alternatives, n to 50.

Four different types of noise, i.e., a white noise, a car horn, a door slam, and a gunshot, were used to corrupt the utterances. These noises each had a wide-band characteristic, for which the sub-band models described earlier had shown no significant merit in comparison to the baseline full-band HMM. These noises were added, respectively, to each of the test utterances to create a partial temporal corruption. The corruption was centered at one of the five positions: beginning, middle, end, a quarter's position and three quarter's position, which was chosen randomly for each test utterance. The duration of the noise was 10% and 20%, respectively, of the duration of the speech utterance. Since the noise only affects a part of the utterance, and since the speech in different parts may have different local energies, it is more appropriate to measure the SNR based on the local average energy rather than on the utterance-level average energy as used earlier. In the following experiments, the SNR was calculated relative to the part of speech where the noise

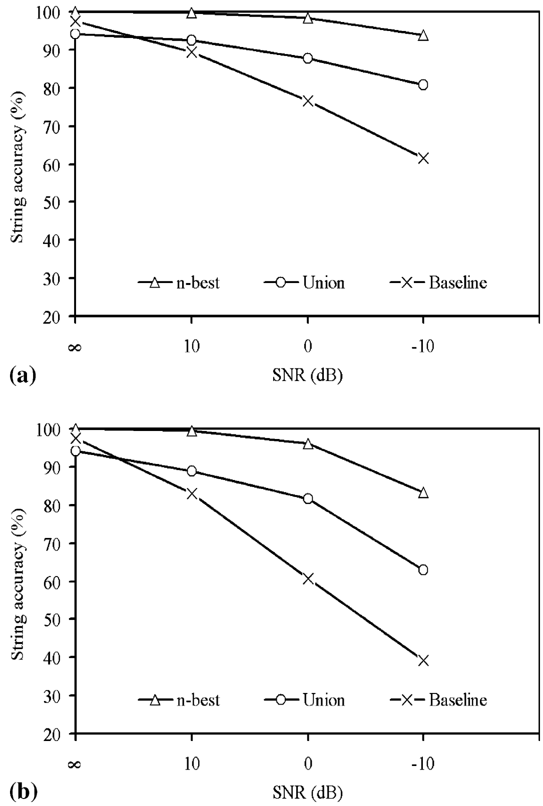


Fig. 3. String accuracy (%) with real-world noise corrupting the duration of each utterance by (a) 10% and (b) 20%, for the segment union model with relative order 0.2 applied to rescore *n*-best alternatives ($n = 50$), and for the baseline HMM.

was added. Fig. 3 presents the average string accuracy obtained by the union model and baseline HMM, respectively, as a function of the SNR. The *n*-best accuracy (i.e., the rate that the correct string is contained in the *n*-best alternatives for rescoring) is also included in the figure. We see that the union model improved upon the baseline model throughout all the noisy conditions. These improvements were achieved at the price of a slight loss of accuracy for clean utterance recognition. Fig. 3 indicates that there is still a large gap between the *n*-best accuracy and the union model accuracy. Note that for a 5-, 6- or 7-digit utterance, a 20% duration corruption may affect the duration of a whole word or longer. This can cause the information of a whole word to be lost, which is difficult to recover without context knowledge. Further improvement may be obtainable by combining with a language model, for recognition of a text sentence.

6. Combination of techniques

So far we have assumed that the noise only causes a partial corruption (1) in the frequency band, (2) in the time duration, or (3) in the feature streams, and we have shown that the union model is capable of dealing with a partial corruption without requiring knowledge about the noise. However, many real-world noises are not partial, i.e., they may exist in all speech frequency bands and accompany the speech throughout its duration. We have studied a combination of the traditional noise-reduction techniques within the union model for dealing with this type of noise. In particular, we assume that the real-world noise may be modelled by a mixture of a known or trainable component and an unknown unexpected component. The trainable component typically includes stationary or slowly varying background noise and/or channel distortion, which may be estimated during non-speech periods thereby being reduced with the conventional noise-reduction techniques, for example, spectral subtraction for additive background noise and cepstral mean subtraction for channel convolutive noise. The unknown unexpected component accounts for the noise that is difficult to estimate, for example, a leftover of an inaccurate noise reduction, or an additional unexpected noise during the utterance, or a combination of these. This component may be dealt with by the union model, if it has a partial corruption characteristic.

As an example, we have built a multi-environment system based on the sub-band union model for speech recognition across a variety of acoustic environments, each involving both an environment-specific noise and some unknown additional noise. The multi-environment technique, by composing an acoustic model for each potential environment, is suited for removing the known or trainable acoustic mismatch across different environments. This technique can be particularly effective if the noise condition in the testing stage remains the same as in the training stage. However, this is usually not the case for real-world applications. The novelty of our system, in contrast to other multi-environment models, is that the acoustic model for each environment is built upon the union model, so that the system is also capable of accommodating further unknown partial corruption within each specific environment. Specifically, we considered the same task for connected digit recognition in four different environments, i.e., clean, car, train, and restaurant (for the last three environments, the SNR for the environmental noise was about 10 dB). The generalized sub-band union model, described in Section 4, was used to build the acoustic model for each environment, which was trained using the noisy training data from the appropriate environment. In recognition, the testing utterances may come from any of the four environments, and they may also be involved in an additional noise corruption which was not expected in the training stage. Three types of band-limited noise were considered as the additional noise: a whistle, a telephone ring, and a ding, as shown in Fig. 2. These were added, respectively, to the test utterances from each testing environment to simulate an additional unknown corruption in that environment. The SNR of the additional noise was 10 dB, so the overall SNR, taking into account both the environmental noise and the additional noise, was about 7 dB. Further, we assumed that for each test utterance, we had no knowledge about the identity of the underlying environment. The final recognition result was selected using the automatic order selection algorithm, described in Section 3, over the models of all the environments. For comparison, we also implemented a multi-environment system based on the same baseline HMM using the full-band features as described in Section 3.

Fig. 4 presents a summary of the performance by the two multi-environment systems based on the union model and baseline model, for recognizing the utterances from each of the testing conditions. As shown, the two systems offered similar accuracy for matched environment training and testing, but the union system offered improved performance for the testing conditions involving additional noise corruption. A further examination of the results by the two systems has

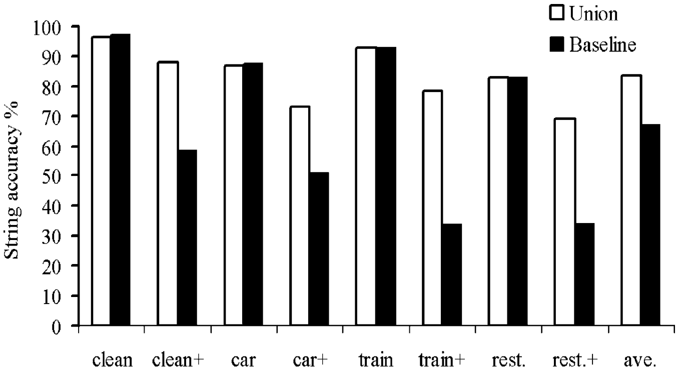


Fig. 4. Summary of the performance of two multi-environment systems based on the union model and baseline model, for each of the test environments (clean, car, train, and restaurant) without additional noise and with unknown additional noise (denoted by "environment+").

indicated that the union model may also be capable of offering improved robustness for an inaccurate noise compensation (Ming, Jancovic, Hanna, & Stewart, 2001). A further example of combining the noise compensation technique with the segment union model described in Section 5 has been studied, for dealing with a mixture of stationary noise and unexpected abrupt noise (Ming, 2001). In principle, there should be no barrier for the traditional techniques to be combined within the union model, to achieve an enhanced capability to deal with the real-world noise.

7. Summary

This paper provided a review of our recent studies on the probabilistic union model and its applications to robust speech recognition. We have shown that the union model may be incorporated into the conventional HMM framework, to provide an enhanced modeling capability for dealing with unknown partial corruptions in the frequency-band, in the time duration, or in the feature streams. We have also studied the combination of the union model with the traditional noise-reduction techniques, for dealing with the mixtures of known corruption and unknown partial corruption.

Currently we are working towards a union model in which the sub-band method and the segment method are combined, for dealing with arbitrary partial corruption within the time-frequency space. A further improvement could be the combination of the union model, which assumes no knowledge about the corruption, with the missing feature method which assumes certain knowledge on the corruption. This combination should reduce the dependence of the missing feature method on the accuracy of the corruption identification, and at the same time, benefit the union model with more certainty on the corrupted features.

Acknowledgements

The authors thank the reviewers for their helpful comments. This work was supported by the UK EPSRC Grant GR/M93734.

References

Baker, J., Cooke, M., Green, P., 2001. Robust ASR based on clean speech models: an evaluation of missing data techniques for connected digit recognition in noise. In: Proceedings of European Conference on Speech Communication and Technology '2001, Scandinavia, pp. 213–216.

Bourlard, H., Dupont, S., 1996. A new ASR approach based on independent processing and recombination of partial frequency bands. In: Proceedings of International Conference on Spoken Language Processing '96, Philadelphia, pp. 426–429.

Cerisara, C., Haton, J.-P., Mari, J.-F., Fohr, D., 1998. A recombination model for multi-band speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing '98, Seattle, pp. 717–720.

Christensen, H., Lindberg, B., Andersen, O., 2000. Employing heterogeneous information in a multi-stream framework. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing 2000, Istanbul, pp. 1571–1574.

Cooke, M., Morris, A., Green, P., 1997. Missing data techniques for robust speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing '97, Munich, pp. 803–806.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication 34, 267–285.

de Veth, J., Cranen, B., de Wet, F., Boves, L., 1999. Acoustic pre-processing for optimal effectivity of missing feature theory. In: Proceedings of European Conference on Speech Communication and Technology '99, Budapest, pp. 65–68.

Daoudi, K., Fohr, D., Antoine, C., 2000. A new approach for multi-band speech recognition based on probabilistic graphical models. In: Proceedings of International Conference on Spoken Language Processing 2000, Beijing.

Drygajlo, A., El-Maliki, M., 1998. Speaker verification in noisy environment with combined spectral subtraction and missing data theory. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing '98, Seattle, pp. 121–124.

Dupont, S., 1998. Missing data reconstruction for robust automatic speech recognition in the framework of hybrid HMM/ANN systems. In: Proceedings of International Conference on Spoken Language Processing '98, Sydney, pp. 1439–1442.

Dupont, S., Bourlard, H., 1997. Using multiple time scales in a multi-stream speech recognition system. In: Proceedings of European Conference on Speech Communication and Technology '97, Rhodes, pp. 3–6.

Hariharan, R., Kiss, I., Viikki, O., Tian, J., 2000. Multi-resolution front-end for noise robust speech recognition. In: Proceedings of International Conference on Spoken Language Processing 2000, Beijing.

Hermansky, H., Tibrewala, S., Pavel, M., 1996. Towards ASR on partially corrupted speech. In: Proceedings of International Conference on Spoken Language Processing '96, Philadelphia, pp. 462–465.

Jancovic, P., Ming, J., 2001a. A probabilistic union model with automatic order selection for noisy speech recognition. Journal of Acoustic Society of America 110, 1641–1648.

Jancovic, P., Ming, J., 2001b. A multi-band approach based on the probabilistic union model and frequency-filtering features for robust speech recognition. In: Proceedings of European Conference on Speech Communication and Technology '01, Aalborg, pp. 1111–1114.

Josifovski, L., Cooke, M., Green, P., Vizinho, A., 1999. State based imputation of missing data for robust speech recognition and speech enhancement. In: Proceedings of European Conference on Speech Communication and Technology '99, Budapest, pp. 2837–2840.

Leonard, R.G., 1984. A database for speaker-independent digit recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing '84, San Diego, pp. 42.11/1–4.

Lippmann, R.P., Carlson, B.A., 1997. Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise. In: Proceedings of European Conference on Speech Communication and Technology '97, Rhodes, pp. 37–40.

McCourt, P., Vaseghi, S., Narte, N., 1998. Multi-resolution cepstral features for phone recognition across speech sub-bands. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing '98, Seattle, pp. 557–560.

Mirghafori, N., Morgan, N., 1998. Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers. In: Proceedings of International Conference on Spoken Language Processing '98, Sydney, pp. 743–747.

Morris, A., Hagen, A., Bourlard, H., 1999. The full-combination sub-bands approach to noise robust HMM/ANN based ASR. In: Proceedings of European Conference on Speech Communication and Technology '99, Budapest, pp. 599–602.

Ming, J., 2001. An improved union model for continuous speech recognition with partial duration corruption. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, Madonna di Campiglio, Trento.

Ming, J., Jancovic, P., Hanna, P., Stewart, D., Smith, F.J., 2000. Robust feature selection using probabilistic models. In: Proceedings of International Conference on Spoken Language Processing 2000, Beijing, pp. 546–549.

Ming, J., Jancovic, P., Hanna, P., Stewart, D., 2001. Modeling the mixtures of known noise and unknown unexpected noise for robust speech recognition. In: Proceedings of European Conference on Speech Communication and Technology '01, Aalborg, pp. 579–582.

Ming, J., Smith, F.J., 1999. Union: a new approach for combining sub-band observations for noisy speech recognition. In: Proceedings of Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere, pp. 175–178.

Ming, J., Smith, F.J., 2000. A probabilistic union model for sub-band based robust speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing 2000, Istanbul, pp. 1787–1790.

Ming, J., Smith, F.J., 2001. Union: a new approach for combining sub-band observations for noisy speech recognition. Speech Communication 34, 41–55.

Ming, J., Stewart, D., Hanna, P., Smith, F.J., 1999. A probabilistic union model for partial and temporal corruption of speech. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, CO, pp. 43–46.

Nadeu, C., Hernando, J., Gorricho, M., 1995. On the decorrelation of the filter-band energies in speech recognition. In: Proceedings of European Conference on Speech Communication and Technology '95, Madrid, pp. 1381–1384.

Okawa, S., Eorico, B., Potamianos, A., 1998. Multi-band speech recognition in noisy environments. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing '98, Seattle, pp. 641–644.

Raj, B., Singh, R., Stern, R.M., 1998. Inference of missing spectrographic features for robust speech recognition. In: Proceedings of International Conference on Spoken Language Processing '98, Sydney, pp. 1491–1494.

Renevey, P., Drygajlo, A., 2000. Statistical estimation of unreliable features for robust speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing 2000, Istanbul, pp. 1731–1734.

Seltzer, M.L., Raj, B., Stern, R.M., 2000. Classifier-based mask estimate for missing feature method of robust speech recognition. In: Proceedings of International Conference on Spoken Language Processing 2000, Beijing, China.

Vizinho, A., Green, P., Cooke, M., Josifovski, L., 1999. Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: an integrated study. In: Proceedings of European Conference on Speech Communication and Technology '99, Budapest, Hungary, pp. 2407–2410.

Wu, S., Kingsbury, B., Morgan, N., 1998. Performance improvements through combining phone- and syllable-length information in automatic speech recognition. In: Proceedings of International Conference on Spoken Language Processing '98, Sydney, pp. 854–857.