ELSEVIER

Speech Communication 34 (2001) 41–55

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Union: A new approach for combining sub-band observations for noisy speech recognition

Ji Ming *, F. Jack Smith

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

Abstract

Recent studies have shown that the sub-band based speech recognition approach has the potential of improving upon the conventional, full-band based model against frequency-selective noise. A critical issue towards exploiting this potential is the choice of the method for combining the sub-band observations. This paper introduces a new method, namely, the probabilistic-union model, for this combination. The new model is based on the probability theory for the union of random events, and represents a new method for modeling partially corrupted observations given little knowledge about the corruption. The new model has been incorporated into a hidden Markov model (HMM) and tested for recognizing a speaker-independent E-set, corrupted by various types of additive noise. The results show that the new model offers robustness to partial frequency corruption, requiring little or no knowledge about the noise statistics. © 2001 Elsevier Science B.V. All rights reserved.

Résumé

Des études récentes ont montré que les techniques de reconnaissance de la parole en sous-bandes peuvent améliorer les performances des systèmes classiques larges bandes dans des conditions de bruit additif à bande étroite. Un aspect important de l'implémentation de l'approche en sous-bandes est le choix de la mèthode de recombinaison des observations dans chaque bande. Ce papier présente une nouvelle méthode, appelée modèle d'union probabiliste, pour effectuer cette recombinaison. Ce nouveau modèle est basé sur la théorie de l'union des événements aléatoires, et représente une nouvelle méthode pour la modélisation d'observations partiellement bruitées sans (ou presque) connaissance a priori de la nature du bruit. Le nouveau modèle a été intégré aux modèles de Markov cachés (HMM) et testé dans une táche de reconnaissance indépendante du locuteur de parole corrompue par divers types de bruits additifs. Les résultats montrent que le nouveau modèle offre une bonne robustesse aux bruits bandes étroites, tout en ne nécessitant pas ou presque pas de connaissance des propriétés statistiques du bruit. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Robust speech recognition; Band-limited noise; Unknown, time-varying noise statistics

1. Introduction

Sub-band based speech recognition is built upon a division of the entire speech frequency-band into several sub-bands, with each sub-band being allocated an acoustic model (e.g. an HMM). Then the likelihoods produced by the individual sub-band

* Corresponding author. Tel.: +44-28-90274723; fax: +44-28-90683890.

E-mail address: j.ming@qub.ac.uk (J. Ming).

PII: S0167-6393(00)00045-5

models are combined to generate an overall likelihood for a given utterance. Recent studies (Bourlard and Dupont, 1996, 1997; Hermansky et al., 1996; Tibrewala and Hermansky, 1997a,b; Cerisara et al., 1998; Mirghafori and Morgan, 1998; Okawa et al., 1998; Bourlard, 1999; Morris et al., 1999) have shown that the sub-band based approach has the potential of producing improved robustness over the conventional, full-band based method against frequency-selective noise. Because this is localized in certain areas of the frequency-band, it affects only some of the sub-bands; the other sub-bands remain unaffected by the noise and can thus be used for recognition. The obvious difficulty is the formulation of the combination of the sub-bands, to exploit this potential. Ideally those bands unaffected or only affected slightly by noise should be selected, as they provide correct information about the utterance, whilst the bands dominated by noise should be excluded as they can be detrimental to the recognition accuracy. This is difficult to achieve without prior information about the noise.

Previous studies have suggested several solutions to this sub-band combination problem. Given the individual likelihoods of the sub-bands, these methods seek different ways of combining these to produce the overall likelihood required for recognition. Typical examples of these combination methods include (see, for example, Bourlard and Dupont, 1996, 1997; Hermansky et al., 1996; Tibrewala and Hermansky, 1997a,b; Cerisara et al., 1998; Okawa et al., 1998):

1. *Product*: where the overall likelihood is defined as the product of the individual sub-band likelihoods, i.e. $\prod_{n=1}^{N} p(o_n)$, where $p(o_n)$ represents the likelihood for the nth sub-band observation o_n, $n = 1, \ldots, N$.
2. *Weighted average*: where the overall likelihood is formed as either a geometric average, i.e. $\prod_{n=1}^{N} p(o_n)^{w_n}$, or an arithmetic average, i.e. $\sum_{n=1}^{N} w_n p(o_n)$, over the individual sub-band likelihoods, where w_n represents the weight for the nth sub-band in the combination.
3. *Neural net*: where neural networks are used to combine the sub-band likelihoods to produce the overall likelihood.

The product method is straightforward but prone to sub-band corruption as the overall likelihood is directly affected by the noisy sub-bands. This method becomes trivial if there is one or more sub-band likelihood turning out to be zero – leading to a zero overall likelihood. Unfortunately, extremely low likelihoods indicate severe violations of the modeling assumption and typically happen for the observations corrupted by noise. This problem may be overcome by weighting the contribution of each sub-band in proportion to their respective signal-to-noise ratio (SNR), as the weighted-average method. Typically, the weighting coefficients can be set to unity for clean or high SNR sub-bands and to zero for low SNR sub-bands. However, this requires knowledge about the noise-band position and the associated SNRs. This knowledge may not be available for some applications involving abrupt noise (e.g. a door bell or a random channel tone) or noise of a non-stationary nature (e.g. a passing car, a telephone ring or a siren). Given that these noises may occur in the middle of a speech utterance and that their characteristics may vary during the utterance, an accurate estimation of the noise statistics may prove impractical, if not impossible.

In the neural-net based approach (e.g. Tibrewala and Hermansky, 1997a), independent networks are trained to estimate the probabilities of all possible combinations of sub-sets of the sub-bands, assuming that there exists at least one combination that accounts for the clean speech. This method faces the problem of how to select the best combination from all the combinations given no knowledge about the noisy bands. Some heuristic methods, such as majority voting or distance pruning, have been studied for this purpose (Hermansky et al., 1996; Tibrewala and Hermansky, 1997a). This problem has been further studied in a more recent model called the full-combination model (Bourlard, 1999; Morris et al., 1999). In this model, the probabilities of different combinations of different sub-bands are combined using a weighted-average method, with each weight representing the relative reliability of a specific set of sub-bands. While it is possible to estimate this reliability in the case of stationary noise, it is generally difficult to estimate this in the case of non-stationary noise with, for example, an unknown, time-varying band position.

In this paper, we present a new approach to the above sub-band combination problem. Specifically, we deal with speech recognition subjected to partial frequency corruption, assuming no knowledge about the band position and statistical distribution of the noise. We call our new method the *probabilistic union model*, which characterizes this partial, unknown corruption based on the union of random events. A distinguishing characteristic of the new model is that it does not require any information about the noise-band position, and as such it is capable of dealing with noise with unknown or time-varying band positions. Knowledge about the noise bandwidth, if available, can be incorporated into the new model to maximize the performance for noises with known or limited bandwidth, but with unknown or time-varying band positions. The new model has been implemented within an HMM framework and tested for speech recognition subject to various types of frequency-selective noise, with unknown or time-varying statistics. The results have shown significant performance achievements for the new model. The work presented in this paper is an extension of a previous conference paper (Ming and Smith, 1999) in both the theoretical and experimental aspects.

This paper is organized as follows. In Section 2 we describe the new model. In Section 3, we discuss the implementation of this model within an HMM framework. The experimental results are presented in Section 4 and a summary is given in Section 5.

2. Probabilistic union model

Assume an N-band system, in which a speech utterance is represented by a set of N sub-band feature streams $\{o_1, o_2, \ldots, o_N\}$, where o_n represents the feature stream from the nth sub-band. The presence of a frequency-selective noise can cause some of the o_n's to be corrupted. Thus, we deal with speech recognition given that some of the sub-band observations o_n's may be noisy, but without appropriate knowledge about the noisy bands, particularly the band position and statistical distribution of the noise.

The product-based combination method, described in Section 1, combines the sub-band observations using the "and" (i.e. conjunction) operator $\wedge$ (although this is not usually explicitly stated), i.e.

$$o_\wedge = o_1 \wedge o_2 \wedge \cdots \wedge o_N, \tag{1}$$

where $o_\wedge$ represents the combined observation. Assuming that the sub-band observations are independent of one another, then the likelihood of $o_\wedge$ equals the product of the individual likelihoods $p(o_n)$'s, i.e.

$$\begin{aligned} p(o_\wedge) &= p(o_1 \wedge o_2 \wedge \cdots \wedge o_N) \\ &= p(o_1)p(o_2)\cdots p(o_N). \end{aligned} \tag{2}$$

When the models, i.e. the probability density functions of the individual sub-bands, $p(x_n)$'s, are trained on clean speech and used for modeling an utterance with some noisy sub-bands, then the corresponding $p(\tilde{o}_n)$'s for the noisy $\tilde{o}_n$'s will be highly inaccurate – when the noise is strong they can become almost zero. This destroys the model's ability to discriminate between correct and incorrect word classes. Simply removing the $p(o_n)$'s with small values from the combination may not improves this, because low likelihoods can also be an indication of a wrong word being assumed for the given utterance. Unless the noisy sub-bands can be identified this is difficult to correct. To solve this problem, we suggest a new approach for representing the observation information. This is the model described below. We start to describe the model assuming no knowledge about the noise bandwidth (except that the noise is *partial* in the frequency domain); then we move to an extension of this model assuming that there is certain knowledge about the noise bandwidth.

2.1. General union model

Given no knowledge about the noisy sub-bands, we can alternatively assume that, in a given set of sub-band observations $o_1, o_2, \ldots, o_N$, the useful features that characterize the speech utterance may be *any* of the o_n's, $n = 1, \ldots, N$, or *any* of the combinations among the o_n's up to the

complete feature set. This can be expressed, using the inclusive "or" (i.e., disjunction) operator $\vee$, as

$$o_{\vee} = o_1 \vee o_2 \vee \cdots \vee o_N, \tag{3}$$

where $o_{\vee}$ is a combined observation based on $\vee$, representing the useful features within $\{o_1, o_2, \ldots, o_N\}$. For example, using a 3-band model, the expression $o_{\vee} = o_1 \vee o_2 \vee o_3$ based on inclusive "or" assumes that the useful features within the given $\{o_1, o_2, o_3\}$ may be o_1, or o_2, or o_3, or $o_1 \wedge o_2$, or $o_1 \wedge o_3$, or $o_2 \wedge o_3$, or $o_1 \wedge o_2 \wedge o_3$. These feature combinations can characterize, respectively, a speech utterance in which there are two-band, one-band and no band corruption, therefore covering all possible partial corruptions, including the no corruption case which may be encountered in a 3-band system. In general, if an observation consists of N components $o_1, o_2, \ldots, o_N$, and these components may be subjected to some partial corruption with unknown characteristics (i.e. coverage, position and intensity of the noise), then the useful information contained in the observation may be modeled by Eq. (3). This model takes into account all possible partial corruptions, thereby requiring no knowledge about the actual noise.

If we assume that the o_n's are discrete random events, then $o_{\vee}$ is the union of the o_n's. Thus, we can compute the probability $P(o_{\vee})$ based on the rules of probability for the union of random events. This probability, for each modeled word, can then be used to decide the recognized word based on the maximum-likelihood principle. Specifically, note that $\vee_{n=1}^{m} o_n = (\vee_{n=1}^{m-1} o_n) \vee o_m$, so $P(o_{\vee})$ can be computed using a recursion (Harris, 1966),

$$P\big(\vee_{n=1}^{m} o_n\big) = P\big(\vee_{n=1}^{m-1} o_n\big) + P(o_m) - P\Big(\big(\vee_{n=1}^{m-1} o_n\big) \wedge o_m\Big) \tag{4}$$

for $m = 2, \ldots, N$. Assuming that the o_n's are mutually independent, then Eq. (4) can be simplified as

$$P\big(\vee_{n=1}^{m} o_n\big) = P\big(\vee_{n=1}^{m-1} o_n\big) + P(o_m) - P\big(\vee_{n=1}^{m-1} o_n\big)P(o_m). \tag{5}$$

Using Eq. (5), the probability of the union of discrete, independent observations can be readily calculated. This computation requires only the probability distributions of the individual sub-bands, i.e. $P(x_n)$'s, which are assumed to be estimated from clean speech training data. In particular, we call Eqs. (3)–(5) the *probabilistic union model*, as opposed to the conjunction-based model Eqs. (1) and (2), which combines observations based on the intersection of events.

Eq. (4) or (5) applies only to probabilities, not to probability densities or likelihoods. While we may estimate probabilities directly for discrete variables, as in the discrete-observation HMMs, we usually obtain only estimates of probability densities for continuous variables. To apply the above union model to continuous random variables (e.g. the sub-band spectral parameters of speech), two methods may be considered. Either, we transform the continuous observations into discrete variables by using vector-quantization techniques, as in the discrete-observation HMMs; or alternatively, we derive an approximate probability for the continuous observation based on the associated probability density. The latter is the method adopted in this paper. Specifically, for each given continuous observation $o_n \in R^K$, we consider the probability of a continuous observation x_n falling into a sub-space surrounding o_n, i.e. $P(x_n \in O_n)$, where O_n is a neighboring sub-space of o_n, defined by

$$O_n = \{x : o_n(k) - \Delta_n(k) \leqslant x(k) \leqslant o_n(k) + \Delta_n(k), \quad k = 1, \ldots, K\}, \tag{6}$$

where $o_n(k)$ (and $x(k)$) represents the kth element of o_n (and x), and $\Delta_n(k)$'s are some positive numbers. This probability is then used to replace $P(o_n)$ in Eq. (5). For small $\Delta_n(k)$'s, $P(x_n \in O_n)$ can be approximated by

$$P(x_n \in O_n) \simeq p(o_n)\Delta_n(1)\Delta_n(2)\cdots\Delta_n(K) = p(o_n)\Delta_n, \tag{7}$$

where $p(x_n)$ is the probability density function for the observation in sub-band n, with $p(o_n)$ giving the likelihood of o_n, and $\Delta_n = \Delta_n(1)\Delta_n(2)\cdots\Delta_n(K)$, equaling the volume of the K-dimensional sub-space O_n. This has been implemented for the union models to deal with continuous observations and will be discussed further in Section 3.

2.2. Extended union model

In this section, we explain, intuitively, the principle of the above union model for sub-band based speech recognition. Then we describe an extension to the model by combining the "and", "or" operators within the model.

Since the $P(o_n)$'s are generally not large, Eq. (5) is effectively the sum of the individual probabilities. The advantage of Eq. (5) over Eq. (2) (i.e. the product model) for noisy speech is that, for $P(x_n)$'s trained for clean speech, the value of $P(\tilde{o}_n)$ for a noisy $\tilde{o}_n$ can be very small and as such makes a small contribution to Eq. (5). Therefore the almost random variation of $P(\tilde{o}_n)$ between the correct and incorrect word classes will have little effect on $P(o_\vee)$. So $P(o_\vee)$ is dominated by noiseless sub-bands. As long as there is one noiseless sub-band and the bandwidth is not too small, Eq. (5) should be able to model corrupted speech with more success than the product model.

The disadvantage of Eq. (5) is that it effectively averages the ability of each sub-band to discriminate between correct and incorrect words, unlike the product model in which each sub-band reinforces the other as the joint probability of the sub-bands is modeled. Since a sub-band may not be able to provide sufficient discriminative information, Eq. (5) is not effective for clean speech.

As a solution, we can combine the use of "and", "or" operators across the sub-bands. For example, in the simple case with four sub-bands, we can have four possible combinations:

1. $o_1o_2o_3o_4$,
2. $o_1o_2o_3 \vee o_1o_2o_4 \vee o_1o_3o_4 \vee o_2o_3o_4$,
3. $o_1o_2 \vee o_1o_3 \vee o_1o_4 \vee o_2o_3 \vee o_2o_4 \vee o_3o_4$,
4. $o_1 \vee o_2 \vee o_3 \vee o_4$,

where the $\wedge$ operator between the o_n's has been omitted. Cases 1 and 4 correspond to the product model (Eq. (1)) and the union model (Eq. (3)), respectively. Cases 2 and 3 are examples of the models in which the $\wedge$ and $\vee$ operators are combined. These are best suited to the situations where there are one or two noisy sub-bands, respectively. For example, in case 2 if there is only one noisy sub-band, the union of the four conjunctions will include one conjunction providing the joint probability of three clean sub-bands, which captures all information of the clean bands. The other three conjunctions each contain a noisy sub-band, with a correspondingly low probability, and therefore make only a small contribution to the union. In a similar way, if there are two noisy sub-bands one of the conjunctions in case 3 will correspond to the two clean sub-bands. This conjunction captures all information of the clean bands and will dominate that union.

In general, if there is a knowledge of the noise bandwidth (not the noise-band position), then this can be incorporated into the union model by combining the use of $\vee$ and $\wedge$ operators. Specifically, we make an assumption about the maximum possible bandwidth (MPB) of the noise. For convenience, we measure this bandwidth in terms of the number of sub-bands that the noise can cover. Therefore, for an N-band system, if the MPB of the noise is M bands ($M < N$), then there exists at least one conjunction of $(N - M)$ bands which characterizes clean speech. Without knowing where the noise occurs, this conjunction may be any of the conjunctions of $(N - M)$ bands. This uncertainty is then represented by the $\vee$ operator. Combining the two together we obtain a model representing clean speech

$$o_\vee = \bigvee_{n_1 n_2 \ldots n_{N-M}} o_{n_1} o_{n_2} \cdots o_{n_{N-M}}, \tag{8}$$

where the "or" is taken over all possible combinations of $n_1, n_2, \ldots, n_{N-M}$ with each $n_i \in (1, \ldots, N)$, giving a total of ${}^{N}C_{N-M}$ combinations. [1] In particular, we call Eq. (8) a *union model of order M*. This model is best suited to the situation where the number of noisy sub-bands equals M, in which case Eq. (8) will include a conjunction of all the remaining $(N - M)$ clean bands for discrimination. However, Eq. (8) may also be suited to situations where the number of noisy sub-bands is less than M. In such cases, Eq. (8) may include more than one conjunction of $(N - M)$ clean bands, each

[1] Since ${}^{N}C_{N-M}$ equals 3 for $N = 4$ and $M = 1$, and 6 for $N = 4$ and $M = 2$, respectively, as shown in cases 2 and 3 above, the number of conjunctions $o_{n_1} o_{n_2} \cdots o_{n_{N-M}}$ which have to be combined in the model Eq. (8) is not large if the number of sub-bands, N, is kept reasonably small.

corresponding to a sub-set of the total clean bands. [2] This may not capture as much discriminative information as the above bandwidth-order matched model, due to the disjunction of clean sub-bands. Eq. (8) is reduced to Eq. (3) when $M = N - 1$. In other words, Eq. (3) is a model capable of accommodating noise with an MPB up to $(N - 1)$ sub-bands. Also, it can be shown that the product model, Eq. (1), is a special case of Eq. (8) with order $M = 0$.

The above indicates that, given an order, the union model defined in Eq. (8) does not make any assumption about the noise-band position. This model, thus, is capable of dealing with noises with unknown or time-varying band positions. It requires a knowledge of the number, or the maximum possible number (i.e., MPB), of corrected bands, for the selection of an appropriate model order. This may be easier to estimate than the exact position of corrupted bands, which actually includes *both* number *and* position of the noisy bands, as required in some traditional sub-band combination methods. Alternatively, we may look at the problem of selecting the model order from a different angle, i.e. from the tolerance of the model towards the loss of frequency information. This means that we can select a model order to accommodate as much noise as possible, subject to a minimum requirement for the recognition performance. This is the method used in the experiments and will be discussed in more detail in Section 4.

The expression for the probability of Eq. (8) can be readily derived with o_n in Eq. (4) replaced by the appropriate conjunctions of sub-bands, i.e. $o_{n_1} o_{n_2} \cdots o_{n_{N-M}}$. Assuming independence between the sub-band observations, then this probability can be computed based on Eq. (5) in two steps:

1. Compute the joint probability $P(o_{n_1} o_{n_2} \cdots o_{n_{N-M}})$ for each of the $(N-M)$-band conjunctions. This equals the product of the individual sub-band probabilities, i.e. $P(o_{n_1})P(o_{n_2}) \cdots P(o_{N-M})$.
2. Compute the probability of the union based on the recursion Eq. (5), with $P(o_n)$ replaced by an appropriate $P(o_{n_1} o_{n_2} \cdots o_{n_{N-M}})$ obtained above.

For two $o_{n_1} o_{n_2} \cdots o_{n_{N-M}}$'s with an overlap, the idempotence property of events is applied when computing the probability of their conjunction. For example, $P(o_1 o_2 o_3 \wedge o_1 o_2 o_4)$ is equivalent to $P(o_1 o_2 o_3 o_4)$. This is applied to modify the last (i.e. product) term of Eq. (5) when conjunctions of overlapped $o_{n_1} o_{n_2} \cdots o_{n_{N-M}}$'s are involved. The above computation requires only the probability distributions of the individual sub-bands, as was required in the general union model discussed in the previous subsection.

3. Incorporation into HMMs

We have built the above union model, Eq. (8), into an HMM for combining the sub-band observations at the *frame* level. The system assumes that the observations are continuous random variables. In the following we first define the model, and then describe the algorithms for model estimation and recognition.

3.1. HMM with union-based sub-band observations

Assume that there are N sub-bands, and that the speech in each sub-band is represented by a sequence of frame vectors $o_n(1), o_n(2), \ldots, o_n(T)$, $n = 1, \ldots, N$. Consider a union-based, frame-level combination of these sub-band observation sequences. This combination is performed within an HMM, by using an observation probability distribution, in state i, of a form $B_i(o_\vee)$, where $o_\vee$ is defined by Eq. (8), with each o_n corresponding to a frame from each of the sub-bands. Denote by $o_\vee(t)$ the combined observation for frames $o_1(t)$, $o_2(t), \ldots, o_N(t)$ at time t. Then the HMM can be expressed as

$$P(o|\lambda) = \sum_s \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} B_{s_t}(o_\vee(t)), \tag{9}$$

where o represents the frame sequences for all the sub-bands, $s = s_0 s_1 \cdots s_T$ is the state sequence of

[2] For example, in case 3 ($M = 2$, $N = 4$) above if there is only one noisy band, o_1, then the model includes three conjunctions of two bands, $o_2 o_3$, $o_2 o_4$, $o_3 o_4$, which characterize clean speech. Each of these conjunctions corresponds to a sub-set of the clean bands $\{o_2, o_3, o_4\}$.

the observation, and λ represents the model parameter set including $\{\pi_i\}$, $\{a_{ij}\}$ and $\{B_i(o_\vee)\}$. As discussed in Section 2, if we assume that the frames across the sub-bands are statistically independent, then the probability $B_i(o_\vee)$ is only a function of the individual frame probabilities $B_i(o_n)$'s. To calculate the probability $B_i(o_n)$ where o_n is a continuous vector, we define a random event $x_n \in O_n$, where O_n is defined by Eq. (6), and then compute the probability $B_i(x_n \in O_n)$ for $B_i(o_n)$. Based on Eq. (7), $B_i(x_n \in O_n)$ can be approximated by $b_i(o_n)\Delta_n$, where $b_i(x_n)$ is the probability density function for the frame vector in sub-band n and state i, and Δ_n represents the volume of the sub-space O_n. Given these, the observation distribution set of the model, $\{B_i(o_\vee)\}$, can thus be represented by the observation density set $\{b_i(x_n)\}$.

3.2. Algorithm for model estimation

Estimating the parameters for the model Eq. (9) is straightforward, if we assume that the model is trained on clean speech data (this is the usual choice given no knowledge about the operating environments). While the union includes all selective combinations among the sub-bands, there is only one combination that best matches the clean speech – the conjunction of all the sub-bands, modeling an observation with no band corruption. Therefore in the training stage we can compute the union-based observation probability, $B_i(o_\vee)$, as $B_i(o_1 \wedge o_2 \wedge \cdots \wedge o_N)$. A more rigorous derivation can also be obtained from the model Eq. (8), which assumes an observation with a maximum of M noisy sub-bands. For clean speech we know that $M = 0$, so we have

$$B_i(o_\vee)\,|_{M=0} = B_i(o_1 o_2 \cdots o_N) = \prod_{n=1}^{N} B_i(o_n). \tag{10}$$

Substituting Eq. (10) into Eq. (9) and replacing each $B_i(o_n)$ with $b_i(o_n)\Delta_n$ for continuous o_n's, we obtain a model characterizing a training utterance

$$\begin{aligned} P(o|\lambda) &\simeq \sum_s \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \prod_{n=1}^{N} b_{s_t}(o_n(t))\Delta_n(t) \\ &= p(o|\lambda) \prod_{t=1}^{T} \prod_{n=1}^{N} \Delta_n(t) \\ &\propto p(o|\lambda), \end{aligned} \tag{11}$$

where $p(o|\lambda)$ is a likelihood function, defined by

$$p(o|\lambda) = \sum_s \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \prod_{n=1}^{N} b_{s_t}(o_n(t)). \tag{12}$$

Based on Eq. (11), we can obtain an estimate of λ by maximizing $p(o|\lambda)$; the $\Delta_n(t)$'s, as scaling factors, are assumed to have no effect on this maximization. This maximization can be accomplished by using the standard Baum–Welch re-estimation algorithm.

3.3. Algorithm for recognition

In recognition, decisions are made by comparing the probabilities, $P(o|\lambda)$'s, between different words. As with the conventional HMMs, we can compute these probabilities by using the Viterbi algorithm, assuming that the most-likely state sequences dominate. This algorithm is shown below,

$$\begin{aligned} \delta_t(j) = \max_i \left(\delta_{t-1}(i) + \log a_{ij}\right) \\ + \log B_j(o_\vee(t)), \end{aligned} \tag{13}$$

where $\delta_t(i)$ is the log probability associated with a best state-sequence ending in state i accounting for the observation sequence up to time t. This algorithm requires the computation of the union-based observation probability, $B_i(o_\vee)$, which in recognition may have an order $M \neq 0$, to account for a certain noisy condition. Using the algorithm described in Section 2, this probability can be computed based on the individual sub-band probabilities $B_i(o_n)$ or $B_i(x_n \in O_n)$; the latter is for continuous observations and is approximated by $b_i(o_n)\Delta_n$. Unlike Eq. (10), the probability $B_i(o_\vee)$ with $M \neq 0$ involves a complex combination of the $b_i(o_n)\Delta_n$'s, such that the effects of the Δ_n's cannot be simply canceled. One way to overcome this problem is to leave out the product term in Eq. (5), assuming that it is small and can be neglected in

comparison to the other two additive terms. Further assume that the Δ_n's are equal (denoted by Δ) for all the sub-bands. Then, based on Eq. (8), the union probability $B_i(o_\vee)$ can be written as

$$\begin{aligned} B_i(o_\vee) &\simeq \sum_{n_1 n_2 \cdots n_{N-M}} B_i(x_{n_1} \in O_{n_1}) B_i(x_{n_2} \in O_{n_2}) \cdots \\ &\qquad B_i(x_{n_{N-M}} \in O_{n_{N-M}}) \\ &\simeq \sum_{n_1 n_2 \cdots n_{N-M}} b_i(o_{n_1}) \Delta b_i(o_{n_2}) \Delta \cdots b_i(o_{n_{N-M}}) \Delta \\ &= \sum_{n_1 n_2 \cdots n_{N-M}} b_i(o_{n_1}) b_i(o_{n_2}) \cdots b_i(o_{n_{N-M}}) \Delta^{N-M} \\ &\propto \sum_{n_1 n_2 \cdots n_{N-M}} b_i(o_{n_1}) b_i(o_{n_2}) \cdots b_i(o_{n_{N-M}}). \end{aligned} \tag{14}$$

Thus, the Δ's become scaling factors which take no effect in recognition based on the maximum-likelihood principle.

Alternatively, a sigmoid function may be used to approximate the probability $B_i(x_n \in O_n)$ based on the likelihood $b_i(o_n)$, i.e.

$$B_i(x_n \in O_n) \simeq \frac{1}{1 + e^{-\ln b_i(o_n)}}. \tag{15}$$

This is equivalent to the assumption that $\Delta_n = 1/(1 + b_i(o_n))$ in terms of the expression $B_i(x_n \in O_n) \simeq b_i(o_n)\Delta_n$. One advantage of Eq. (15) is that it produces an approximated probability that is proportional to the likelihood value, and at the same time satisfies the constraint $0 \leqslant B_i(x_n \in O_n) < 1$ (this is required by Eq. (5) such that it will not produce a negative probability). The probability $B_i(o_\vee)$ with each $B_i(x_n \in O_n)$ defined by Eq. (15) can thus be computed based on Eq. (5), including the product term. Because this term is usually very small indeed (particularly for models with an $M \ll N$), the two methods described above based on Eqs. (14) and (15) have been found to produce very close results.

4. Experiments

A number of databases have been used to test the new union model. These include the Connex speaker-independent alphabetic database provided by British Telecom (Woodland and Cole, 1991; Ming and Smith, 1996), and the TIDIGITS database (Leonard, 1984). All these experiments have shown similar performance achievements for the new model. In this paper we focus on the experiments with the Connex alphabetic database, from which the E-set words (b, c, d, e, g, p, t, v) were extracted for the tests.

This database contains three repetitions of each word by 104 speakers, 53 male and 51 female. Among the 104 speakers, 52 have been designated for training and the other 52 for testing; these were both roughly balanced with respect to both sex and age. For each word, then, about 156 utterances are available for training, and a total of 1219 utterances are available for testing for all eight words.

For this database, a previous baseline HMM, based on a diagonal-covariance matrix structure and a feature vector of 25 elements (12 MFCCs + 12ΔMFCCs + 1Δ Energy) for each frame, has achieved an accuracy of 85.7 (Woodland and Cole, 1991). A similar result was obtained by Ming and Smith (1996) based on a similar model configuration. Although later more sophisticated models have been applied to this database, resulting in improved accuracies (see, for example, Woodland and Cole, 1991; Valtchev, 1995; Ming and Smith, 1996; Hanna et al., 1999; Ming et al., 1999a), these were not implemented in this paper as optimizing the baseline model is not the main theme of this study.

The speech is divided into frames of 25.6 ms, with a between-frame overlap of 15.6 ms. For each frame, a multi-channel, Mel-scaled filter bank is used to estimate the log-amplitude spectra of the speech. These filter-bank channels are then grouped uniformly into sub-bands, for which features are extracted for being used as the sub-band based observations. In particular, the models with 3, 5 and 7 sub-bands, respectively, are implemented, which are grouped from the filter banks with 35 channels (for the models with 5 and 7 sub-bands) and 36 channels (for the model with 3 sub-bands), respectively. For each frame of each sub-band, a feature vector, consisting of MFCCs and their first-order delta parameters, is calculated. The size of this sub-band frame vector is balanced among the three models, so that the overall frame vector (consisting of all the sub-band frame

vectors) in each model can have a similar size. For example, in our experiments we chose a sub-band frame vector of 14 elements for the 3-band model, of 8 elements for the 5-band model, and of 6 elements for the 7-band model. Within each vector, half of the elements are for MFCCs and the other half for the delta parameters. A 15-state HMM is estimated for each word, with the last nine states being tied among all the eight words (Woodland and Cole, 1991). For comparison, we also implemented a full-band model using the same state structure and feature vectors as used by Woodland and Cole (1991), and a product model described in Section 1. All models are based on Gaussian densities with diagonal covariance matrices.

4.1. Tests with clean speech

Firstly, we test the new union model for recognizing clean speech. We examine different model configurations and determine the suitable model structures for dealing with speech recognition involving partial frequency corruption.

Based on Eq. (8), for each model with N sub-bands, recognition can be performed with different model orders (i.e. M) from 0 to $N-1$. Table 1 presents the recognition results, with the number of bands $N=3$, 5 and 7, respectively. Table 1 also includes the respective results produced by the product model and full-band model.

As described earlier, since there is no band corruption, a clean speech utterance is better characterized by the conjunction of all the sub-band observations. This explains why the product model, derived from such a conjunction, produced the best performance. However, recalling that the product model corresponds to a union model with order $M=0$, the performance of the union model improves rapidly as the order is reduced.

Table 1 indicates that, given a subdivision of the frequency-band, the performance of the union model decreases as the order is increased. This is because a higher order model bases recognition on the conjunction of fewer sub-bands. Table 1 reveals the recognizability of E-set based on partial frequency-bands, and in particular the amount of frequency information needed to achieve a certain accuracy for clean E-set recognition. For example, in the case $N=5$, Table 1 indicates that to obtain a recognition accuracy of above 80%, a model order lower than 4 is needed. This corresponds to the requirement of using at least 2 sub-bands (covering 40% of the entire frequency-band) of each utterance for recognition. Any partial frequency-bands narrower than this may not contain sufficient discriminative information and therefore may not be able to deliver the required performance. In other words, the model may only be able to tolerate a loss of frequency information up to 3 sub-bands, in terms of a recognition accuracy of above 80%. This knowledge, for clean speech, may be used to determine an order for the union model for dealing with both clean and noisy speech (assuming no knowledge about which is actually being processed). Since there is no point to develop a model that is not effective for clean speech, we select the model order to accommodate as much noise as possible subject to a minimum requirement for clean speech recognition performance. Later we will present an example using such a method for selecting the union model order. In that example, we choose an order $M=3$ in a 5-band system; this model should be capable of dealing with any noise which corrupts up to 3

Table 1
Recognition results for clean E-set by the new union model, product model and full-band model

Number of bands N	Union model accuracy (%)							Product model accuracy (%)	Full-band model accuracy (%)
	Model order M								
	0	1	2	3	4	5	6		
3	88.8	85.4	80.2	–	–	–	–	88.8	84.0
5	87.0	86.6	84.7	81.7	76.1	–	–	87.0	
7	87.3	87.1	85.1	82.3	80.0	75.6	70.4	87.3	

sub-bands of each utterance, while providing a clean speech recognition performance of about 81%, based on Table 1.

To effectively isolate any local frequency corruption from the other usable bands, we may subdivide the frequency-band into small sub-bands. However, small sub-bands mean poor ability per band for discrimination, and subsequently poor performance for the system. For example, Table 1 shows that the 3-band model achieved an accuracy of 80.2% at order 2, effectively using single-band information (covering one third of the entire frequency-band). However, for the 7-band model to achieve a similar accuracy (i.e. 80.0%, at order 4), three sub-bands are required, which cover over 42% of the entire frequency-band. This indicates that breaking the available frequency-band into too many sub-bands may cause a loss of discriminative information. This phenomenon has been discovered previously by Hermansky et al. (1996) and Bourlard and Dupont (1996). An optimum balance between the noise localization and linguistic discrimination remains a topic for study. In the following we choose the 5-band system for further experiments on noisy E-set recognition. This model is a compromise between the 3-band and 7-band models. It improves upon the 3-band model with a higher band resolution, and upon the 7-band model with a greater single-band discriminative capability.

4.2. Tests with simulated nullifying noise

Since the presence of noise rarely increases the probability of the correct word, it is of interest to first test the model by setting the probabilities of certain bands to zeros – a simple simulation of the possible effects of some strong noise occurring in the corresponding sub-bands. We chose the bands to be nullified randomly on each utterance basis, and manually corrupted these bands such that they produced approximately zero probabilities over the correct model. This simple test immediately disqualifies the product model since it produces constantly near-zero overall likelihoods for the correct word. Table 2 presents the recognition results produced by the union model and product model. As shown in Table 2, given a nullification condition, the performance of the union model varies with the model order – it is maximized when the order of the model matches the actual number of corrupted bands, and it becomes poor when the number of nullified bands is underestimated. As described in Section 2, the match between the model order and noise bandwidth leads to an inclusion of all the remaining clean bands into a conjunction, thereby capturing all information of the clean bands. However, Table 2 also indicates that, while such a match is desirable to maximize the performance, the model with a higher order provides higher robustness against variations in the number of corrupted bands. Therefore in applications in which unknown or time-varying noises are involved, the model with a higher order may be chosen to provide the required robustness.

4.3. Tests with stationary narrow-band noise

We then test the new union model in the presence of stationary narrow-band noise. The noise, added to the speech, is generated by passing the Gaussian white noise through a band-pass filter. In the experiments, we define the SNR based on the averaged energy of all the test speech utterances.

Table 2
Recognition results by the union model and product model for E-set corrupted by simulated nullifying noise

Number of nullified bands N	Union model accuracy (%)				Product model accuracy (%)
	Model order M				
	1	2	3	4	
1	85.9	84.5	81.1	74.0	14.1
2	11.8	81.3	78.9	73.6	12.4
3	9.6	13.6	75.8	69.4	9.3
4	10.1	13.1	15.4	62.9	9.6

So the noise in each test utterance exhibits a constant loudness, regardless of the actual loudness of speech in that utterance.

We fix the 3-dB cut-off bandwidth of the noise as 100 Hz and vary the central frequency of the noise across the sub-bands. In particular, four central frequencies are chosen, which are 900, 1800, 2700 and 3500 Hz. Tables 3 and 4 present the recognition results with an SNR of 10 and 0 dB, respectively. These two tables also include the respective results produced by the product model and full-band model.

Tables 3 and 4 show that the union model offers a remarkable improvement over both the product and full-band models. The improvement is significant at all the four orders adopted for the union model. Specifically, Table 3 indicates that as compared with the product model, the union model reduced the error rate by an average of 54.4%, 56.6%, 49.2% and 31.4%, respectively, at the orders of 1, 2, 3 and 4. As the SNR drops to 0 dB (Table 4), these corresponding error reductions are increased to 58.2%, 61.6%, 55.0% and 40.1%, respectively. While noises with central frequencies of 900, 1800 and 3500 Hz were mainly located within one sub-band, noise with a central frequency of 2700 Hz was located around the border of two adjacent bands, thereby affecting both bands, and explaining why the union model with order 2 produced the best accuracy. Instead of exploring further comparisons between the union model and other methods (e.g. majority voting or distance pruning, suggested by Hermansky et al. (1996) and Tibrewala and Hermansky (1997a), for neural-net based combination), we conducted a direct comparison with a model based on cheating, which assumes a complete knowledge of the corrupted bands and removes those bands manually from the combination. The results obtained by cheating are shown in Tables 3 and 4 in the last column. As shown, the union model, based only on a rough estimate of the number of corrupted

Table 3
Recognition results for E-set corrupted by stationary narrow-band noise with a bandwidth of 100 Hz and different central frequencies (SNR = 10 dB)

Noise central frequency (Hz)	Union model accuracy (%)				Product model accuracy (%)	Full-band model accuracy (%)	Cheat accuracy (%)
	Model order M						
	1	2	3	4			
900	84.7	80.7	78.3	71.6	50.0	47.9	84.4
1800	83.8	81.5	77.2	69.1	62.5	48.9	83.8
2700	70.1	81.5	78.2	71.7	56.2	51.9	80.0
3500	85.4	83.9	81.4	73.2	64.3	43.7	83.9
Average	81.0	81.9	78.8	71.4	58.3	48.1	83.0

Table 4
Recognition results for E-set corrupted by stationary narrow-band noise with a bandwidth of 100 Hz and different central frequencies (SNR = 0 dB)

Noise central frequency (Hz)	Union model accuracy (%)				Product model accuracy (%)	Full-band model accuracy (%)	Cheat accuracy (%)
	Model order M						
	1	2	3	4			
900	78.2	75.1	72.1	66.3	26.2	31.3	77.8
1800	78.1	75.6	69.8	61.8	39.8	32.3	78.0
2700	62.4	76.5	74.0	64.1	44.6	29.0	76.2
3500	82.0	81.5	77.4	67.5	52.3	22.4	79.9
Average	75.2	77.2	73.3	64.9	40.7	28.8	78.0

Table 5
Recognition results for E-set corrupted by narrow-band noise with a time-varying central frequency

SNR (dB)	Union model accuracy (%)				Product model accuracy (%)	Full-band model accuracy (%)
	Model order M					
	1	2	3	4		
10	79.1	81.2	77.9	72.3	44.1	37.0
0	61.8	72.8	71.9	64.9	17.3	23.9

bands, may achieve equally good performance as the cheating model. The union model has the capability of ignoring those noisy bands that significantly violate the statistics of the training data population; but none of these bands is physically removed from the combination. This may explain why occasionally the union model may obtain slightly better results than the cheating model, due to the remaining contributions of some noisy bands.

4.4. Tests with time-varying narrow-band noise

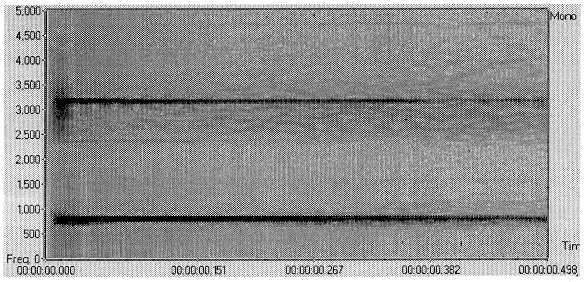
Further experiments were conducted examining the capability of the union model against a type of time-varying narrow-band noise, with a changing central frequency during the utterance. For this type of noise, it is generally impractical to assume the availability of a prior estimate of the noise-band positions, based on which a band weighting can be effected (the weighted-average model, Section 1). We show by a simple experiment that the union model offers robustness to this type of noise. Specifically, each test utterance was corrupted by the same type of narrow-band noise as described above, except that the central frequency of the noise changed during the utterance from 900 to 1800 Hz and then to 2700 Hz, each frequency lasting approximately an equal duration. The recognition results produced by the union model are presented in Table 5, along with the results by the product model and full-band model.

4.5. Tests with real-world noise

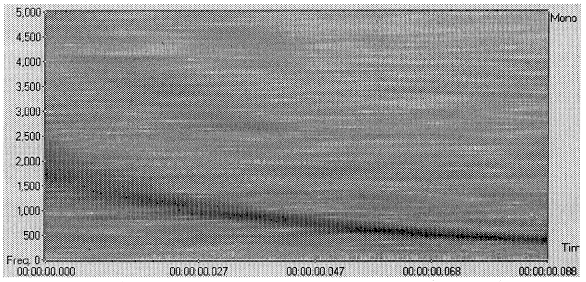
Finally, we tested the new union model for recognizing speech subjected to some real-world noises, which have a dominant frequency-localization characteristic. The noise data used in the experiments are shown in Fig. 1, which include the sounds of a ding, a laser and a telephone ring. [3] As shown in Fig. 1, the ding includes two prominent frequency components, located around 800 Hz and 3200 Hz, respectively; the laser has one dominant narrow-band component, with both the bandwidth and central frequency being time-varying; the telephone ring has four prominent frequency components, located at about 1300 Hz, 1650 Hz, 3400 Hz and 4250 Hz, respectively. In addition to the respective dominant frequency components, each sound also includes many minor, complex frequency components distributed over the entire frequency-band. These noise data were added, respectively, to each of the test utterances for recognition experiments. Table 6 presents the recognition results obtained by the union model, product model and full-band model, respectively, with an SNR of 10 dB.

Table 6 indicates that the union model offers a significant improvement over the other two models, throughout all test conditions. Table 6 also indicates that the improvement by the union model for the telephone-ring noise is less significant in comparison to the improvement for the other two types of noise. This is because the telephone-ring noise has a particular multi-band characteristic: the first two tones lying in the border of bands 2 and 3, and the last two tones falling within band 4, which thus affects about 3 sub-bands. In fact, we have experienced the ineffectiveness of the sub-band based method for dealing with wide-band noise, as experienced by Tibrewala

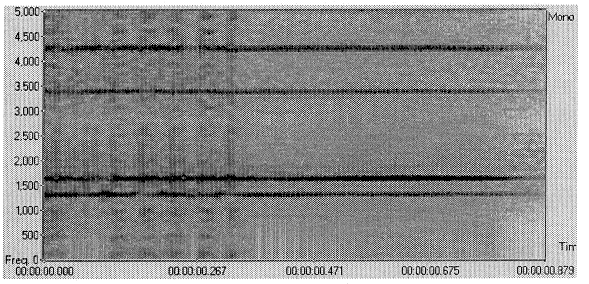
[3] These data were extracted from the sound files "ding.wav", "laser.wav" and "ringin.wav", respectively, provided in the Windows NT operating system.



(a) Ding



(b) Laser



(c) Telephone ring

Fig. 1. Spectrograms of real-world noise data used in the experiments: (a) ding; (b) laser; (c) telephone ring.

Table 6
Recognition results for E-set corrupted by real-world noises (SNR = 10 dB)

Noise type	Union model accuracy (%)				Product model accuracy (%)	Full-band model accuracy (%)
	Model order M					
	1	2	3	4		
Ding	69.2	74.9	79.5	70.1	34.9	40.1
Laser	65.1	71.9	68.3	61.4	36.2	41.8
Telephone ring	54.9	65.1	62.8	59.5	37.8	47.3
Average	63.1	70.6	70.2	63.7	36.3	43.1

and Hermansky (1997a,b), and by Okawa et al. (1998). Wide-band noise affects all sub-bands, which therefore violates the noise-localization assumption made in the sub-band based model. For a system to be capable of dealing with both narrow-band and wide-band noises, a combination of different techniques may be needed. Research on this is currently being undertaken.

4.6. Selection of model order

In the above experiments we ignored the problem of how to select the union model order; instead, we ran the experiments for all possible orders. As discussed earlier, without knowledge about the operating environment (e.g. clean or noisy), this problem may be

Table 7
Summary of performance of the union model with 5 sub-bands and order $M = 3$ in various test conditions and comparisons with other models

Noise condition	SNR (dB)	Union model	Product model	Full-band model	Cheat model
Clean		81.7	87.0	84.0	
Stationary	10	78.8	58.3	48.1	83.0
narrow-band	0	73.3	40.7	28.8	78.0
Time-varying	10	77.9	44.1	37.0	
narrow-band	0	71.9	17.3	23.9	
Ding	10	79.5	34.9	40.1	
Laser	10	68.3	36.2	41.8	
Telephone ring	10	62.8	37.8	47.3	
Average		74.3	44.5	43.9	

solved by choosing the highest order, subject to an acceptable performance for clean speech recognition, i.e. seeking a compromise between the maximum accuracy and robustness. For example, based on the above experimental results, we may build a 5-band system using an order $M = 3$. The performance of this model, in all above test conditions, is summarized in Table 7, along with the performances by the product model and full-band model. The union model reduced the error rate by an average of 53.7% and 54.2%, respectively, in comparison to the other two models.

5. Summary

This paper described a new statistical approach, the probabilistic union model, for combining sub-band observations for speech recognition subjected to partial frequency corruption. This model improves upon the previous sub-band combination methods in that it assumes no knowledge about the noise characteristics, particularly the frequency-band position and statistical distribution of the noise. This lack of knowledge can be experienced, for example, when a noise with an unknown or time-varying nature occurs in the middle of an utterance. The new model characterizes the partially and randomly corrupted observations based on the probability theory for the union of random events. It has been implemented within an HMM framework, for combining the sub-band features at the frame level. Experiments on the recognition of a speaker-independent E-set, corrupted by various types of frequency-selective noise, have shown the great potential of the new model for dealing with unknown, time-varying band-limited noise. This implemented system can be readily applied to continuous speech recognition. Also, the principle of combining features based on the union can be extended to the combination of units at a higher level than frame, e.g. phoneme or syllable (Dupont and Bourlard, 1997). This provides a new means of handling partially corrupted information given little knowledge about the corruption. We believe that this new model has applications in many areas of signal processing and pattern recognition involving partial unknown feature corruption. For example, a recent application of this model to the recognition of speech subjected to partial *temporal* corruption has shown significant performance (Ming et al., 1999b).

Acknowledgements

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

References

Bourlard, H., 1999. Non-stationary multi-channel (multi-stream) processing towards robust and adaptive ASR. In: Proceedings of Workshop on Robust Methods for Speech Recognition in Adverse Conditions. Tampere, Finland, pp. 1–10.

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. Philadelphia, USA, pp. 426–429.

Bourlard, H., Dupont, S., 1997. Sub-band based speech recognition. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing. Munich, Germany, pp. 1251–1254.

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. Seattle, USA, pp. 717–720.

Dupont, S., Bourlard, H., 1997. Using multiple time scales in a multi-stream speech recognition system. In: Proc. Eurospeech. Rhodes, Greece, pp. 3–6.

Hanna, P., Ming, J., Smith, F.F., 1999. Interframe dependence arising from preceding and succeeding frames-application to speech recognition. Speech Communication 28, 301–312.

Harris, B., 1966. Theory of Probability. Addison-Wesley, Reading, MA.

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

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

Ming, J., Smith, F.J., 1996. Modeling of interframe dependence in an HMM using conditional Gaussian mixtures. Comput. Speech Language 10, 229–247.

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, Finland, pp. 175–178.

Ming, J., Hanna, P., Stewart, D., Owens, M., Smith, F.J., 1999a. Improving speech recognition performance by using multi-model approaches. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing. Phoenix, USA, pp. 161–164.

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

Mirghafori, N., Morgan, N., 1998. Transmissions and transitions: a study of two common assumptions in multi-band ASR. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing. Seattle, USA, pp. 713–716.

Morris, A.C., Hagen, A., Bourlard, H., 1999. The full-combination sub-bands approach to noise robust HMM/ANN based ASR. In: Proc. Eurospeech. Budapest, Hungary, pp. 599–602.

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. Seattle, USA, pp. 641–644.

Tibrewala, S., Hermansky, H., 1997a. Sub-band based recognition of noisy speech. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing. Munich, Germany, pp. 1255–1258.

Tibrewala, S., Hermansky, H., 1997b. Multi-band and adaptation approaches to robust speech recognition. In: Proc. Eurospeech 97. Rhodes, Greece, pp. 2619–2622.

Valtchev, V., 1995. Discriminative methods in HMM-based speech recognition. Ph.D. Dissertation, Cambridge University Engineering Department, England.

Woodland, P.C., Cole, D.R., 1991. Optimizing hidden Markov models using discriminative output distributions. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing. Toronto, Canada, pp. 545–548.