A POSTERIOR UNION MODEL FOR IMPROVED ROBUST SPEECH RECOGNITION IN NONSTATIONARY NOISE

Ji Ming and F. Jack Smith

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

ABSTRACT

This paper investigates speech recognition with partial feature corruption, assuming unknown, time-varying noise characteristics. We extend our previous probabilistic union model from a conditional-probability formulation to a posterior-probability formulation. The new formulation allows the order of the model to be optimized for every single frame, and therefore greatly enhances the capability of the model for dealing with nonstationary noise corruption. Experiments have been conducted on two databases: TIDigits with noise corruption and Aurora 2, to demonstrate the improved robustness for the new model. Examples are presented showing that the new model can co-exist with existing noise-reduction techniques to provide improved noise robustness.

1. INTRODUCTION

A speech recognition system needs to be robust against unknown partial corruption of the acoustic features, where some of the feature components may be corrupted by noise, but knowledge about the corruption, including the number and identities of the corrupted components and the characteristics of the corrupting noise, is not available. This problem has been addressed recently by the missing feature method (see, for example, [1]–[8]), and by the probabilistic union model [9], [10]. The missing feature method proposes a solution to the problem by identifying and ignoring the feature components that are strongly affected by the noise and thus carry no reliable information about the utterance. The key problem is how to determine which components are corrupt when no knowledge about the noise is assumed. A number of methods have been suggested for identifying the corruption, for example, based on an estimate of the noise characteristics obtained during speech pauses [3]–[6], or based on a knowledge of the speech such as the harmonicity from auditory scene analysis [7], [8]. The probabilistic union model represents an alternative, aiming to lift the requirement for identifying the noisy feature components. Assuming a feature set comprising N components, M of which are corrupt, the union model deals with the uncertainty of the corrupted components by performing a disjunction to combine every $(N-M)$ sized subset of the components, assuming that any one of the subsets may be the set which contains all the clean components and no others, giving reliable information about the utterance. This effectively reduces the problem of identifying the noisy components to a problem of estimating the number of the noisy components, i.e., M. This number defines the order of the union model.

Previously we studied the formulation of the union model with conditional probabilities. A major drawback of this formulation is the difficulty in estimating the order, particularly for corruptions with a nonstationary nature. In an application of the model for combining the short-term subband features for speech recognition, a heuristic method was studied for estimating the order based on the state-duration counts from a hidden Markov model (HMM) [11], [12]. This method assumed a constant order for a whole utterance and thus gave only a suboptimal performance in nonstationary noise conditions.

In this paper, we extend the union model from the conditional probability formulation to a posterior probability formulation, to overcome the above problem. The new formulation allows the order to be optimized for every single frame based on the maximum *a posteriori* (MAP) rule, thereby greatly enhancing the capability of the model for dealing with nonstationary noise corruption.

2. PROBLEM FORMULATION

Let $X=(x_1,x_2,...,x_N)$ be a feature set consisting of N components, where x_n represents the nth feature component. In speech recognition, for example, X may be a frame feature vector consisting of N components from different feature streams. A recognizer's job is to correctly classify each X into one of the K classes, C_1, C_2, ..., C_K, representing different speech units. The classification may be based on the conditional probability $P(X \mid C_k)$ or on the *a posteriori* probability $P(C_k \mid X)$. Assume that some of the feature components x_n in X are noisy, but without knowledge about their identity. We term this unknown partial feature corruption. A partial feature corruption may be caused by the noise that affects only some of the feature streams, for example, a band-limited noise affecting only certain parts of the speech frequency band, and a convolutional noise (i.e. channel effect) affecting the static cepstra more adversely than the delta cepstra. In addition, a partial feature corruption may also be the result of inaccurate noise-reduction processing, which leaves some of the components distorted due to insufficient knowledge of the noise.

Assume that in X there are M noisy components with an unknown identity. Denote by X_{N-M} the subset in X which contains the $(N-M)$ clean components. The probabilistic union model deals with the uncertainty of X_{N-M} by using the "or" (i.e. disjunction) operator to combine every $(N-M)$ sized subset of the components, assuming that *any* one of the subsets may be X_{N-M}. The conditional probability of X_{N-M} based on a union model can be written as [9], [10]

$$P(X_{N-M} \mid C_k) = P(\bigvee_{n_1 n_2 \cdots n_{N-M}} x_{n_1} x_{n_2} \cdots x_{n_{N-M}} \mid C_k) \quad (1)$$

This work was supported by the UK EPSRC grant GR/M93734.

where $\vee$ denotes the "or" operator, $x_{n_1}x_{n_2}\cdots x_{n_{N-M}}$ is a subset in X containing $(N-M)$ components as a probable candidate for X_{N-M}, and the "or" operator $\vee$ is applied between all possible subsets of $(N-M)$ components in X. The parameter M in (1) defines the order of the model, which corresponds to the number of corrupted components in X. For partial corruption, M assumes a value within the range $0 \leq M \leq N-1$, accommodating from no component corruption up to $N-1$ component corruption in X.

We call (1) the conditional-probability union model. The expression for $P(X_{N-M} \mid C_k)$ is readily derived using the rules of probability for the union of random events. Assuming independence between the feature components, $P(X_{N-M} \mid C_k)$ can be written as [9], [10]

$$P(X_{N-M} \mid C_k) \simeq \sum_{n_1 n_2 \cdots n_{N-M}} P(x_{n_1} \mid C_k)P(x_{n_2} \mid C_k)\cdots P(x_{n_{N-M}} \mid C_k) \tag{2}$$

where $P(x_n \mid C_k)$ is the conditional probability of the component x_n, and the summation is over all possible subsets of $(N-M)$ components taken from X. Since (2) is the sum of the individual subset probabilities, its value is dominated by the subset probabilities with large values. Therefore, if we can assume that the clean-component subset produces a large probability for the correct class, then selecting the maximum value of $P(X_{N-M} \mid C_k)$ with respect to C_k has a chance to get the correct class C_k for X without requiring the identity of the M noisy components. However, when the value of M is unknown, the classification can not be performed based on the maximum value of $P(X_{N-M} \mid C_k)$ with respect to M and C_k. This is because, for a specific C_k, the values of $P(X_{N-M} \mid C_k)$ for different M are of a different order of magnitude and are thus not directly comparable. For unknown or time-varying noisy environments, estimating the order M for (2) can be a difficult task. In the following we present a new formulation for the union model which overcomes this problem.

3. THE POSTERIOR UNION MODEL

3.1. The model

Consider the problem of classifying an N-component feature set $X = (x_1, x_2, ..., x_N)$ into one of the K classes C_1, C_2, ..., C_K, assuming that there are M ($0 \leq M < N$) components in X being corrupted by noise, but neither the value of M nor the identity of the corrupted components are known *a priori*. We deal with this problem based on the *a posteriori* union probability. Let X_{N-M} denote the subset in X containing the $(N-M)$ clean components, the *a posteriori* union probability of class C_k given X_{N-M} is defined as

$$P(C_k \mid X_{N-M}) = \frac{P(X_{N-M} \mid C_k)P(C_k)}{\sum_{j=1}^{K} P(X_{N-M} \mid C_j)P(C_j)} \tag{3}$$

where $P(X_{N-M} \mid C_k)$ is the conditional union probability of order M as defined in (1) and (2), and $P(C_k)$ is the class prior which is assumed not to be a function of the order M.

With a constant class prior, (3) is similar to (2) in that it is dominated by the subset probabilities with large values. Therefore, if we assume that the clean subset produces a large probability for the correct class, selecting the maximum $P(C_k \mid X_{N-M})$ is likely to get the correct class C_k for X without requiring the identity of the M noisy components. A major difference between (3) and (2) is that the *a posteriori* union probability $P(C_k \mid X_{N-M})$ is normalized for the orders. Therefore we can obtain an optimal estimate for the unknown order M for each class based on the maximum *a posteriori* (MAP) rule, i.e.

$$\hat{M} = \arg\max_{M} P(C_k \mid X_{N-M}) \tag{4}$$

This leads to an optimal classifier that implements a joint MAP decision for order estimation and feature classification:

$$X \in C_k \text{ if } P(C_k \mid X_{N-\hat{M}}) = \max_{j} \max_{M} P(C_j \mid X_{N-M}) \tag{5}$$

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

3.2. Incorporation into a hidden Markov model (HMM)

The above posterior union model has been incorporated into an HMM for combining the short-term subband features with unknown band-limited corruption. Let $X(t) = (x_1(t), x_2(t), ..., x_N(t))$ be a short-time measurement (i.e. frame) at time t consisting of N independent subband feature streams, with $x_n(t)$ being the feature stream from the nth subband. Consider the classification of each frame $X(t)$ into an HMM state s_t, $s_t \in \{1, 2, ..., K\}$, based on the *a posteriori* union probability $P(s_t \mid X_{N-M_t}(t))$ of order M_t, which is defined, based on (3), as

$$P(s_t \mid X_{N-M_t}(t)) = \frac{P(X_{N-M_t}(t) \mid s_t)P(s_t)}{\sum_{\nu_t=1}^{K} P(X_{N-M_t}(t) \mid \nu_t)P(\nu_t)} \tag{6}$$

where $P(s_t)$ is the state prior and $P(X_{N-M_t}(t) \mid s_t)$ is the conditional union probability of order M_t in state s_t, which can be written, based on (2), as

$$P(X_{N-M_t}(t) \mid s_t) \simeq \sum_{n_1 \cdots n_{N-M_t}} P(x_{n_1}(t) \mid s_t)\cdots P(x_{n_{N-M_t}}(t) \mid s_t) \tag{7}$$

where $P(x_n \mid i)$ is the emission probability for subband feature stream x_n in state i, and the summation is over all possible subsets of $(N-M_t)$ subband streams taken from $X(t)$.

To apply (6) to an HMM, we first express the traditional HMM in terms of the *a posteriori* probabilities of the states. Denote by $X_1^T = (X(1), X(2), ..., X(T))$ a speech utterance of T frames and by $S_1^T = (s_1, s_2, ..., s_T)$ a state sequence for X_1^T. The joint probability of X_1^T and S_1^T based on an HMM is defined as

$$\begin{aligned} P(X_1^T, S_1^T \mid \lambda) &= \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} P(X(t) \mid s_t) \\ &= \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \frac{P(X(t) \mid s_t)}{P(X(t))} P(X(t)) \\ &= \pi_{s_0} \prod_{t=1}^{T} \frac{a_{s_{t-1}s_t}}{P(s_t)} P(s_t \mid X(t)) \prod_{t=1}^{T} P(X(t)) \end{aligned} \tag{8}$$

where $P(s_t \mid X(t))$ is the *a posteriori* probability of state s_t given $X(t)$, and $P(s_t)$ is the state prior. The last product, $\prod_{t=1}^{T} P(X(t))$, is not a function of the state index and, thus, has no effect in recognition. Replacing $P(s_t \mid X(t))$ in (8) with the *a posteriori* union

probability $P(s_t \mid X_{N-M_t}(t))$ defined in (6), we thus have a new model for the joint probability of X_1^T and S_1^T:

$$P(X_1^T, S_1^T \mid \lambda, M_1^T) \propto \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} P(s_t \mid X_{N-M_t}(t)) \quad (9)$$

In (9), we have assumed a constant state prior, $P(s_t)$, for simplicity; $M_1^T = (M_1, M_2, ..., M_T)$ is the sequence of the orders for the individual frames in X_1^T. Recognition is performed with the following modified Viterbi algorithm that includes the MAP decision for optimizing the order for each frame:

$$\delta_t(j) = \max_i [\delta_{t-1}(i) a_{ij}] \max_{M_t} P(s_t = j \mid X_{N-M_t}(t)) \quad (10)$$

where $\delta_t(i)$ represents the probability of the best state path up to time t that ends in state i.

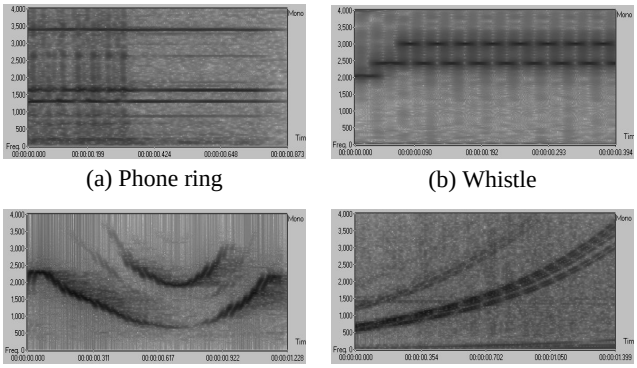
4. EXPERIMENTAL RESULTS

4.1. Experiments on TIDigits

The database contained 6196 test utterances for connected-digit recognition. For comparison, we used the same feature vector as used in [12] for each frame. The speech was sampled at 8 kHz and segmented into frames of 256 samples. Each frame was divided into five subbands, and each subband was modeled by three static MFCCs plus three delta MFCCs. So we had a total of 10 features streams (five for MFCCs and five for ΔMFCCs) for each frame. Each digit was modeled by a left-to-right HMM with ten states, and each state consisted of eight Gaussian mixtures with diagonal covariance matrices. This paper is focused on the comparison between the new posterior union model and the conditional union model, described in [12] and above. For a comparison between the conditional union model and other methods for subband combination, see [12]. The conditional union model assumed a constant order for a whole utterance and selected the order by comparing the HMM state sequences associated with different orders.

Fig. 1 shows the real-world noises used in the test, including a telephone ring, a whistle, and the sounds of "contact" and "connect", extracted from an Internet tool. These noises each had a dominant band-selective nature, and the noises "contact" and "connect" were particularly nonstationary. These noises were added, respectively, to each of the test utterances with different levels of signal-to-noise ratio (SNR). Table I shows the average digit-string accuracy over the four noise conditions, obtained by the new posterior union model, by the conditional union model, and by a baseline HMM using ten full-band MFCCs plus delta MFCCs for each frame. The posterior union model improved upon the conditional union model throughout all test conditions. These improvements are due to the frame-level optimization for the order selection implemented in the new model. The conditional union model used a constant order for all frames, and its performance was thus compromised by the time-varying noise characteristics.

Improved performance was also obtained for the new model in stationary band-limited noise. Table II shows the results averaged over eight different stationary noise conditions, including three cases with one subband corruption, three cases with two subband corruption, and two cases with three subband corruption, within the five subbands of the system [12]. The frame-level order optimization enables the new model to extract an optimal number of feature components from each frame (this number may be varying from frame to frame due to the varying frame-level SNR), thereby obtaining more information for correct recognition than the conditional union model.



(a) Phone ring (b) Whistle

(c) Contact (d) Connect

Fig. 1. Real-world noises used in the experiments

Table I. Average digit string accuracy in real-world noise, for the new posterior union model, compared to the conditional union model and a baseline HMM

SNR (dB)	Posterior Union	Conditional Union	Baseline HMM
Clean	96.42	96.21	97.53
10	87.96	85.07	51.47
5	81.13	78.43	26.74
0	71.05	68.63	11.20

4.2. Experiments on Aurora 2

To increase the band resolution for the Aurora noise, we increased the number of subbands from five to twelve, by using 12 decorrelated filter-bank energies [13], with a decorrelation filter $H(z) = 1 - z^{-1}$, plus the delta and delta-delta parameters, as the feature set for each frame. The feature set for each frame thus contained 36 parameters, representing 36 different feature streams. We present the results for Test Set A based on clean-condition training.

Table III presents the word accuracy obtained by the new posterior union model, including the relative improvement in word error rate when compared to the ETSI baseline system using a full-band MFCC front-end [14]. The new model improved over the baseline system throughout all noise conditions.

The union model can be added onto other noise-reduction techniques to provide improved robustness against the inaccuracy in noise reduction. To achieve best performance, the conventional

Table II. Average digit string accuracy in stationary band-limited noise

SNR (dB)	Posterior Union	Conditional Union	Baseline HMM
10	92.45	89.90	52.99
5	89.33	86.33	29.97
0	83.47	80.91	13.93

noise-reduction techniques require certain knowledge such as the spectral or cepstral characteristics of the noise. An accurate estimation of these characteristics can be difficult if the noise is unpredictable and/or nonstationary. The residual noises introduced by an inaccurate noise-reduction processing may be modeled as partial, unknown, time-varying corruption and can be dealt with by the union model. This combination reduces the dependence of the system on the accuracy of the noise estimation. Tables IV and V show an example, in which a Wiener-filtering front-end was employed to enhance the speech utterances before recognition. The noise spectrum used to build the Wiener filter for each utterance was simply estimated using the first ten frames of each utterance without any further adaptation. Table IV shows the results with the use of Wiener filtering alone without the use of the union model, and Table V shows the results when Wiener filtering and the union model were combined. Comparing Table IV with Table III shows that the Wiener filtering operation caused a degradation for the "babble" noise condition. This degradation was effectively avoided by the inclusion of the union model, as shown in Table V. In addition, Table V indicates that the inclusion of the union model improved the accuracy for most noise conditions.

5. CONCLUSIONS

This paper described a new statistical method – the posterior union model, for speech recognition involving partial feature corruption assuming no knowledge about the noise. The new model is an extension of our previous union model from a conditional-probability formulation to a posterior-probability formulation. The experimental results based on the TIDigits and Aurora 2 databases indicate that the new formulation improves the performance of the union model and enhances its capability for modeling nonstationary noise corruption. Examples were presented which show that the new model can be effectively combined with other noise robust techniques to provide improved performance.

Table III. Word accuracy and error reduction (ER) by the posterior union model on test set A, clean training, relative to the ETSI baseline system

SNR	Sub.	Bab.	Car	Exhib.	Ave.	%ER
Clean	98.31	98.52	98.64	98.85	98.58	-30.98
10 dB	78.97	83.56	80.79	75.16	79.62	36.90
5 dB	59.07	63.66	59.35	51.34	58.36	31.20
0 dB	31.19	32.93	29.02	22.89	29.01	14.52
Ave.	66.89	69.67	66.95	62.06	66.39	
%ER	8.16	43.70	28.73	6.48		23.99

Table IV. Results obtained by the use of a Wiener-filtering front-end without the use of the union model

SNR	Sub.	Bab.	Car	Exhib.	Ave.	%ER
Clean	97.97	98.55	98.33	98.24	98.27	-43.35
10 dB	80.01	71.40	87.77	77.48	79.17	35.51
5 dB	68.22	52.96	71.61	56.49	62.32	37.75
0 dB	40.19	26.31	39.96	29.10	33.89	20.39
Ave.	71.60	62.31	74.42	65.33	68.41	
%ER	21.22	30.03	44.84	14.54		28.56

Table V. Results obtained by the combination of Wiener-filtering and the posterior union model

SNR	Sub.	Bab.	Car	Exhib.	Ave.	%ER
Clean	98.22	98.67	98.24	98.77	98.48	-35.52
10 dB	82.84	81.56	87.71	81.24	83.34	48.42
5 dB	68.62	65.27	72.92	60.78	66.89	45.29
0 dB	41.82	35.95	46.53	33.51	39.45	27.09
Ave.	72.88	70.36	76.35	68.58	72.04	
%ER	24.77	44.98	48.99	22.55		36.77

6. REFERENCES

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

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

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

[4] B. Raj, R. Singh, and R. M. Stern, "Inference of missing spectrographic features for robust speech recognition," *ICSLP'98*, pp. 1491-1494.

[5] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, pp. 267-285, 2001.

[6] P. Renevey and A. Drygajlo, "Statistical estimation of unreliable features for robust speech recognition," *ICASSP'2000*, pp. 1731-1734.

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

[8] J. Baker, M. Cooke, and P. Green, "Robust ASR based on clean speech models: an evaluation of missing data techniques for connected digit recognition in noise," *Eurospeech'2001*, pp. 213-216.

[9] J. Ming and F. J. Smith, "Union: a new approach for combining sub-band observations for noisy speech recognition," *Speech Commun.*, vol. 34, pp. 41-55, 2001.

[10] J. Ming and F. J. Smith, "Union: a model for partial temporal corruption of speech," *Comput. Speech Language*, vol. 15, pp. 217-231, 2001.

[11] P. Jancovic and J. Ming, "A probabilistic union model with automatic order selection for noisy speech recognition," *J. Acoust. Soc. Amer.*, vol. 110, pp. 1641-1648, 2001.

[12] J. Ming, P. Jancovic, and F. J. Smith, "Robust speech recognition using probabilistic union models," *IEEE Trans. Speech Audio Processing*, vol. 10, pp.403-414, 2002.

[13] K. K. Paliwal, "Decorrelated and liftered filter-bank energies for robust speech recognition," *Eurospeech'99*, pp. 85-88.

[14] H.-G. Hirsch and D. Pearce, "The Aurora experimental frameowrk for the performance evaluation of speech recognition systems under noisy conditions," *ISCA ITRW ASR2000 "Automatic speech recognition: challenges for the next millennium"*, 2000.