MODELING SUB-BAND CORRELATION FOR NOISE-ROBUST SPEECH RECOGNITION

James McAuley, Ji Ming, Philip Hanna and Darryl Stewart

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

ABSTRACT

This paper investigates the effect of modeling sub-band correlation for noisy speech recognition. Sub-band data streams are assumed to be independent in many sub-band based speech recognition systems. However, the structure and operation of the human vocal tract suggests this assumption is unrealistic. A novel method is proposed to incorporate correlation into sub-band speech feature streams. In this method, all possible combinations of sub-bands are created and each combination is treated as a single frequency band by calculating a single feature vector for it. The resulting feature vectors capture information about every band in the combination as well as the dependency across the bands. Experiments conducted on the TIDigits database demonstrate significantly improved robustness in comparison to an independent sub-band system in the presence of both stationary and non-stationary noise.

1. INTRODUCTION

Recent studies have demonstrated the ability of sub-band speech recognition systems to offer improved robustness compared to conventional full-band models, particularly in the presence of frequency selective noise (e.g., [1]–[4], [7]). This is due to their ability to isolate noise corruption within particular frequency bands by splitting the frequency spectrum into sub-bands. Sub-band speech recognition systems usually extract features from individual sub-bands based on the assumption that sub-band data streams are independent. An examination of the structure and operation of the human vocal tract however suggests this assumption is unrealistic.

Experiments based on independent sub-band features indicate that ignoring correlation between sub-band feature streams causes recognition performance to degrade quite rapidly as the number of sub-bands increases. Figure 1 shows an example of experiments conducted on the TIDigits database, using mel-frequency cepstral coefficients (MFCCs) to model independent sub-bands. As the number of sub-bands increases, the frequency range covered by each sub-band becomes smaller and so provides less discriminative information. The lack of correlation between the smaller sub-bands leads to poor recognition performance.

There may be two possible approaches to modeling correlation between sub-band features. Firstly, correlation may be captured in the acoustic modeling stage by using, for example, full-covariance matrices between sub-band features [3] or full-combination neural nets [2][4]. Secondly, correlation may be captured in the feature extraction stage, by developing a new feature format. An example of this is described in [5] to model correlation in sub-band based speaker recognition. In that study, a highly redundant sub-band architecture was employed whereby, from an overall set of 24 filter-bank channels, every possible combination of 20 consecutive channels was created. A feature vector was calculated for each combination and independent speaker recognition experiments were performed on each of these feature vectors. No attempt was made to fuse the independent decisions from each combination into a single overall decision.

The first author thanks the European Social Fund (ESF) for funding.

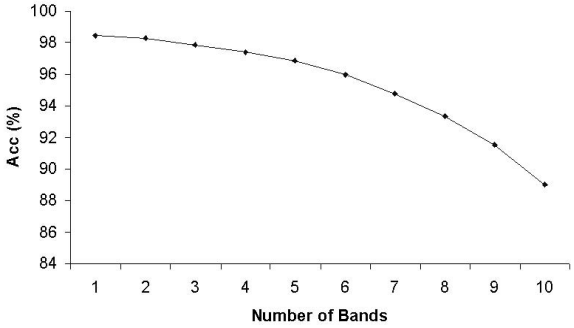


Fig. 1. String accuracy (%) versus number of sub-bands using independent sub-band (ISB) features in clean conditions

In this paper, a new feature extraction approach to modeling sub-band correlation is discussed. To isolate noise, we create all possible combinations between the sub-bands, assuming there is at least one combination containing all clean sub-bands. To capture correlation between the sub-bands, we treat each of these combinations as a single frequency band and calculate a single feature vector for it. The obtained feature vector therefore captures information about every band in the combination as well as the dependency across the bands. It can be shown that the features described in [5] are included as a subset in our new feature set.

2. MODELING SUB-BAND CORRELATION

The proposed method is based on a filter-bank cepstral analysis approach. Traditional full-band speech recognition systems extract feature vectors by applying a Discrete Cosine Transformation (DCT) across all the filter-bank energies (FBEs) together, obtaining the full-band cepstral feature vector. Sub-band systems typically assume independence between sub-bands and extract feature vectors by applying the DCT within separate groups (i.e. sub-bands) of FBEs. This obtains the respective cepstral feature vectors, one for each sub-band, which are assumed to be independent of each other in the acoustic modeling for speech recognition. We call this the independent sub-band system.

This paper proposes to model sub-band correlation by applying the DCT to combinations of sub-bands taken as single bands.

To illustrate this technique let us assume that for each frame of speech there are 8 FBEs ($f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8$) to be grouped uniformly into four sub-bands: Band 1 = ($f_1 f_2$) , Band 2 = ($f_3 f_4$), Band 3 = ($f_5 f_6$) and Band 4 = ($f_7 f_8$) and that there is one band, Band 2 say, which is corrupted by noise. Bands 1, 3 and 4, therefore, correspond to the clean bands that should be utilised for recognition. Instead of processing these three bands independently, they are grouped together into a single band ($f_1 f_2 f_5 f_6 f_7 f_8$) and the DCT is applied to the combined bands taken as a single band. This produces C_{134}, which represents the correlated cepstral feature vector for bands 1, 3 and 4. This vector not only features the individual bands in the combination but also captures the correlation between them. The identity of the corrupt band will not normally be available a priori so we must apply this principle to all possible combinations of three bands, assuming that there is one combination that contains all clean sub-bands. The resulting feature set is denoted by $\{C_{ijk}\}$ for all combinations of i, j and k in the range $1 \leq i, j, k \leq 4$.

In a similar way, let us assume that there are two bands, Bands 1 and 3 say, corrupted by noise. Bands 2 and 4 will therefore correspond to the clean bands that should be utilised for recognition. These two bands are grouped together into a single band ($f_3 f_4 f_7 f_8$) and the DCT is applied. This creates C_{24}, which represents the correlated cepstral feature vector for bands 2 and 4. Again, since the identity of the two corrupt bands will not normally be available a priori, we must apply this principle to all possible combinations of two bands, assuming that there is one combination that contains the remaining two clean sub-bands. The resulting feature set is denoted by $\{C_{ij}\}$ for all combinations of i and j in the range $1 \leq i, j \leq 4$.

In general, to account for M band corruption with unknown identities in a system with N sub-bands, we create all possible combinations of $N - M$ bands and treat each combination as a single band by calculating a single vector for each combination. The resulting feature set is denoted by $\{C_{n_1 \ldots n_{N-M}}\}$, for all combinations of n_i in the range $1 \leq n_i \leq N$. When the number of corrupted bands is unknown, this feature set is created for every possible M from 0 to $N-1$, assuming zero or partial band corruption, and that there is thus one feature vector within these feature sets that corresponds to all the remaining $N-M$ clean bands. The overall feature set containing all these sub-sets of feature vectors can be denoted by $C = \{ C_{n_1}, C_{n_1 n_2}, C_{n_1 n_2 n_3}, \ldots, C_{n_1 \ldots n_N} \}$ for all combinations of n_i in the range $1 \leq n_i \leq N$. This feature set includes the feature vectors for each of the individual sub-bands (as used in the independent sub-band system) and also a feature vector for the complete band (as used in the full-band system), to account for the situations in which there are $N - 1$ corrupt bands and zero corrupt bands respectively. This correlated sub-band system therefore includes both the independent sub-band system and the full-band system as special cases.

3. ACOUSTIC MODELING

Given the feature set that contains feature vectors corresponding to all possible combinations between the sub-bands, the task of the recogniser is to select the feature vector corresponding to all clean sub-bands to use for recognition. In the missing-feature methods, sub-band feature streams usually need to be labeled as reliable or corrupt for their recombination in the final classification decision. A number of techniques have been studied for this purpose, for example, the contribution of each sub-band feature stream to the overall combination decision can be weighted by an estimation of the local signal-to-noise ratio (SNR) in each band [1]. Recently, several studies have attempted to release the need for identification of the corrupted features. These include, for example, the full-combination model [4], the acoustic backing-off model [6], and the posterior union model (PUM) [7][8]. In this paper the PUM is employed to select from the given feature set the feature vector corresponding to the clean or less contaminated sub-bands.

Let $C = \{C_{n_1}, C_{n_1 n_2}, C_{n_1 n_2 n_3}, \ldots, C_{n_1 \ldots n_N}\}$ represent the entire feature set for a given frame in a system consisting of N sub-bands, where, as described earlier, each $C_{n_1 \ldots n_b}$ corresponds to a feature vector capturing correlation for a certain combination of b sub-bands, i.e. bands $n_1, \ldots, n_b$, in the range $1 \leq n_i \leq N$. Assume that there are M bands being corrupted by noise (assuming that $0 \leq M \leq N-1$ for partial frequency-band corruption), and that $X_{n_1 \ldots n_{N-M}} \in C$ is the feature vector modeling the remaining $N - M$ clean sub-bands. The task is then to select from C this clean vector for recognition, assuming no knowledge about its identity. The PUM deals with the uncertainty of $X_{n_1 \ldots n_{N-M}}$ by assuming that it can be *any* of the feature vectors $C_{n_1 \ldots n_{N-M}}$ for $N - M$ sub-bands, i.e., it can be expressed as a union of all possible random vectors $C_{n_1 \ldots n_{N-M}}$. Based on the PUM, the conditional probability of $X_{n_1 \ldots n_{N-M}}$ given a speech state s can be written as

$$\begin{aligned} P(X_{n_1 \ldots n_{N-M}}|s) &= P\Big(\bigvee_{n_1 \ldots n_{N-M}} C_{n_1 \ldots n_{N-M}}|s\Big) \\ &\approx \sum_{n_1 \ldots n_{N-M}} P(C_{n_1 \ldots n_{N-M}}|s) \end{aligned} \quad (1)$$

where $\vee$ denotes the union (i.e. "or") operator, which is applied over all possible feature vectors of $(N - M)$ distinct sub-bands. A posterior union probability of state s given $X_{n_1 \ldots n_{N-M}}$ can be defined as

$$P(s|X_{n_1 \ldots n_{N-M}}) = \frac{P(X_{n_1 \ldots n_{N-M}}|s)P(s)}{\sum_s P(X_{n_1 \ldots n_{N-M}}|s)P(s)} \quad (2)$$

where $P(X_{n_1 \ldots n_{N-M}}|s)$ is the conditional union probability defined in (1) and $P(s)$ is the prior probability for state s.

An optimal estimation of M, i.e. the number of noisy sub-bands, can be obtained based on (2), using the following maximum a posteriori (MAP) rule:

$$\hat{M} = \arg\max_M P(s|X_{n_1 \ldots n_{N-M}}) \quad (3)$$

The optimized posterior union probability, $\max_M P(s|X_{n_1 \ldots n_{N-M}})$, is used to replace the state-based emission probability in an HMM for frame vector $X_{n_1 \ldots n_{N-M}}$ associated with state s. This model requires neither the identity not the number of corrupted sub-bands.

4. EXPERIMENTAL RESULTS

4.1. Conditions

The TIDigits database was used for the experiments. This database contains 6196 test utterances for connected-digit recognition. The speech was sampled at 8 kHz and segmented into frames of 200 samples. Each frame was divided into five sub-bands, and each sub-band was modeled by an equal number of static MFCCs and delta MFCCs. For independent sub-bands 4 MFCCs (i.e. two static and two delta) were used. For combinations of 2, 3, 4 &

Table 1. String accuracy rates and error reduction rate (ERR) for correlated sub-band (CSB) features and independent sub-band (ISB) features in clean conditions

Type	Performance
CSB	98.13
ISB	96.51
ERR	46.42

5 sub-bands, 10, 14, 18 & 22 MFCCs were used respectively. For each combination, the static and delta MFCCs were merged into a single feature vector for a frame. Each digit was modeled by a left-to-right HMM with sixteen states, and each state consisted of three Gaussian mixtures with diagonal covariance matrices. A 3-state HMM was used to account for the silences surrounding each utterance, the middle state of which was used to account for short inter-digit pauses within the connected digit utterances.

Experiments were conducted using the clean test utterances from the TIDigits database and using the same utterances corrupted by various stationary and non-stationary noises. To generate stationary noise, Gaussian white noise was passed through a band-pass filter with a bandwidth of 100Hz at various central frequencies to create the effect of 1, 2 or 3-band corruption within the five sub-bands. The noise was added to the clean test utterances at signal-to-noise ratios of 0, 5 and 10dB, respectively. For non-stationary noise, five different mobile phone ringtones were sampled at 8kHz and added to the clean test utterances at signal-to-noise ratios of 0, 5 and 10dB, respectively.

To study the effect of modeling sub-band correlation on recognition performance, experiments were conducted using independent sub-band (ISB) features and correlated sub-band (CSB) features in ideal conditions, i.e. where the number of corrupted bands is known. This knowledge is used to define the value M for each frame in the PUM. To investigate the potential for modeling correlation in practical conditions, i.e. where the number of corrupted bands is not known, the same experiments were conducted using the PUM where the number of noisy bands, M, was automatically estimated frame-by-frame based on the joint MAP algorithm described in (3).

4.2. The effect of modeling sub-band correlation

Experiments on CSB features and ISB features were conducted in clean conditions and in various simulated and real-world noise environments. The number of noisy bands was assumed known in the PUM such that any observed improvement in performance would be mainly attributable to modeling sub-band correlation. In these conditions, the number of noisy bands is either known a priori (for stationary noise) or selected based on best performance (for non-stationary noise).

Table 1 shows a 46.42% reduction in errors is achieved by modeling sub-band correlation for clean speech. Table 2 shows the average string accuracy rates and error reduction rates (ERR) for experiments conducted in various stationary narrow-band noise conditions. Improvements in performance by modeling correlation are observed in all noise conditions, particularly in the presence of 2-band noise corruption. Greater ERRs are achieved at higher signal-to-noise ratios. Figure 2 shows the spectra of the mobile phone ringtones used as real-world noise. It demonstrates that the noises are non-stationary and have a dominant band-selective nature. The results in Table 3 show that modeling correlation reduces recognition errors for all types of noise.

Table 2. Average string accuracy and ERR for CSB features and ISB features in various band-selective stationary noise conditions

		Number of corrupt bands		
SNR (dB)	Type	1 band	2 band	3 band
10	CSB	94.90	93.87	81.77
	ISB	90.47	85.54	73.35
	ERR	46.54	57.60	31.58
5	CSB	92.88	92.67	76.87
	ISB	87.73	83.62	70.21
	ERR	41.98	55.22	22.34
0	CSB	87.84	90.19	70.68
	ISB	82.69	80.54	65.25
	ERR	29.77	49.57	15.64

Table 3. String accuracy and ERR for CSB features and ISB features in real world non-stationary noise conditions

		Mobile Phone Ringtone Type				
SNR (dB)	Type	1	2	3	4	5
10	CSB	95.24	89.74	90.45	92.21	93.02
	ISB	93.21	81.15	80.62	89.69	86.23
	ERR	29.90	45.57	50.72	24.44	49.31
5	CSB	94.68	87.07	87.79	91.58	90.99
	ISB	93.11	77.39	75.76	88.15	83.93
	ERR	22.79	42.81	49.63	28.95	43.93
0	CSB	93.48	84.44	83.43	90.03	88.42
	ISB	84.25	64.79	62.82	79.11	70.06
	ERR	58.60	55.81	55.43	52.27	61.32

Modeling sub-band correlation therefore improves recognition performance in clean conditions and in all tested stationary and non-stationary noise conditions.

4.3. Modeling sub-band correlation based on PUM

Tables 1–3 in the previous section indicate that capturing sub-band correlation improves the performance, assuming that the number of noisy sub-bands is known. Tables 4–6 in this section show the results obtained by the PUM assuming no knowledge (i.e. identity and number) of the noisy sub-bands. The number of noisy sub-bands is estimated for each frame based on the MAP decision in (3), for both CSB and ISB features.

Table 4 shows a 31.35% reduction in errors is achieved for clean speech by modeling correlation between sub-bands. The

Table 4. String accuracy rates and ERR for CSB and ISB features in clean conditions using the PUM based on MAP optimization

Type	Performance
CSB	95.38
ISB	93.27
ERR	31.35

PUM therefore achieves a comparable ERR by modeling correlation for clean speech to that obtained in ideal conditions (46.42%), as shown in Table 1. Table 5 shows reductions in recognition errors are observed in all tested narrow-band stationary noise conditions using correlated features, particularly in 2-band noise corruption. In 3-band noise corruption, the PUM achieves a very similar performance to that obtained in ideal conditions, as shown in Table 2. Table 6 shows modeling correlation also reduces recognition errors in all tested real-world non-stationary noise conditions using the PUM. For non-stationary noise, the PUM based on MAP optimization performs almost as well as in ideal conditions (as shown in Table 3) because it estimates the number of noisy sub-bands on a frame-by-frame basis.

The results in Tables 4–6 show that the PUM based on MAP estimation of the number of noisy bands is able to perform almost as well as, and sometimes better than, the PUM assuming that the number of noisy bands is known, i.e. it is able to capture most of the benefits of modeling correlation without requiring knowledge of the noisy bands.

Table 5. Average string accuracy rates and ERRs for CSB features and ISB features in various band-selective stationary noise conditions using the PUM based on MAP optimization

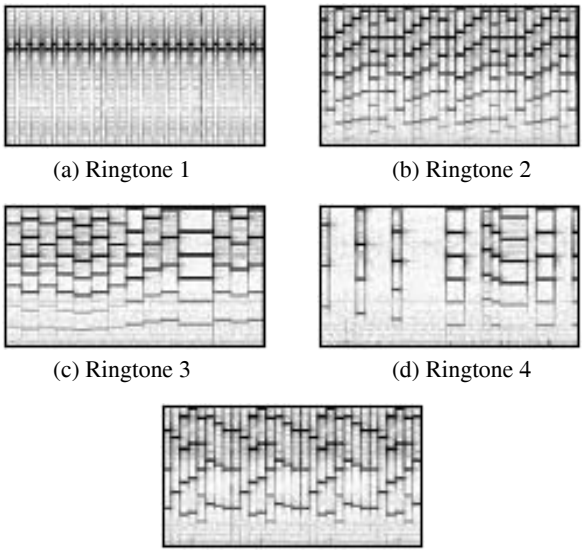
		Number of corrupt bands		
SNR (dB)	Type	1 band	2 band	3 band
10	CSB	91.15	89.01	81.02
	ISB	87.41	82.37	73.16
	ERR	29.72	37.67	29.27
5	CSB	87.92	86.41	74.88
	ISB	83.87	79.37	68.08
	ERR	25.07	34.14	21.30
0	CSB	81.23	82.45	67.78
	ISB	78.14	75.78	62.28
	ERR	14.11	27.54	14.57

Table 6. String accuracy and ERR for CSB features and ISB features in real world non-stationary noise conditions using the PUM based on MAP optimization

		Mobile Phone Ringtone Type				
SNR (dB)	Type	1	2	3	4	5
10	CSB	94.85	89.29	87.84	93.27	91.93
	ISB	92.23	82.77	76.53	89.79	86.96
	ERR	33.72	37.84	48.19	34.08	38.11
5	CSB	94.06	86.98	84.88	92.62	90.22
	ISB	91.85	78.39	70.93	88.40	83.94
	ERR	27.12	39.75	47.99	36.38	39.10
0	CSB	91.80	83.78	81.25	91.17	87.12
	ISB	90.56	71.65	63.10	86.02	80.12
	ERR	13.14	42.79	49.19	36.84	35.21

5. CONCLUSIONS

This paper proposed a new feature format for sub-band speech recognition. The new feature format captures correlation between sub-bands and improves robustness in the presence of both stationary and non-stationary band-selective noises and in noise-free conditions. Correlated sub-band features incorporate useful information that independently extracted sub-band features disregard. Furthermore it is shown that the posterior union model based on MAP estimation for the number of noisy bands is able to capture most of the benefits of modeling correlation without requiring knowledge about the noisy bands.



(a) Ringtone 1 (b) Ringtone 2

(c) Ringtone 3 (d) Ringtone 4

(e) Ringtone 5

Fig. 2. Frequency spectra of the real-world noises used in tests

6. REFERENCES

[1] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," *ICSLP'96*, pp.426–429.

[2] H. Hermansky, M. Pavel and S. Tibrewala, "Towards ASR on partially corrupted speech," *ICSLP'96*, pp.462–465.

[3] S. Okawa, E. Bocchieri and A. Potamianos, "Multi-band speech recognition in noisy environments," *ICASSP'98*, pp.641–644.

[4] A. Morris, A. Hagen and H. Bourlard, "The full combination sub-bands approach to noise robust HMM/ANN based ASR," *Eurospeech'99*, pp.599–602.

[5] L. Besacier, J.F. Bonastre, "Subband approach for automatic-speaker recognition," *European Journal Signal Processing*, vol. 80, pp.1245–1259, 2000.

[6] J. de Veth, B. Cranen, and L. Boves, "Acoustic backing-off in the local distance computation for robust automatic speech recognition," *ICSLP'98*, pp.65-68.

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

[8] J. Ming and F.J. Smith, "A Posterior Union Model For Improved Robust Speech Recognition In Nonstationary Noise," *ICASSP'03*, pp.420–423.