Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Information Sciences 156 (2003) 21–38

INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL

www.elsevier.com/locate/ins

Reduced feature-set based parallel CHMM speech recognition systems

Waleed H. Abdulla [a,*], Nikola Kasabov [b]

[a] *Electrical and Electronic Engineering Department, School of Engineering, The University of Auckland, Private Bag 92019, Auckland, New Zealand*
[b] *Knowledge Engineering and Discovery Research Institute (KEDRI), Auckland University of Technology, Private Bag 92006, Auckland, New Zealand*

Received 30 December 2000; received in revised form 20 June 2001; accepted 10 March 2003

Abstract

This paper presents the multi-streams paradigm as a technique for improving speech signal feature set design and as a performance booster for speech recognition systems, based on the continuous-density hidden Markov model (CHMM) framework. In the multi-streams paradigm we are dealing with different feature sets independently to estimate the same task, and then combining their results at a suitable stage. This paradigm combines the strengths of many varied feature vectors to attain better statistical estimation. Under the proposed paradigm the feature vectors are split into three independent streams, and each stream is used to model an independent CHMM. Then the outcomes of these models, when subjected to any speech input, are merged under a certain strategy. This technique alleviates the dominance effect of the features, and reduces the dimensionality of the feature vectors used in each model. The F-ratio technique is used to further reduce the dimensionality of each stream. Experimental results on different datasets show superiority of the developed paradigm over the corresponding single-stream baseline.

Keywords: Speech recognition; Hidden Markov modelling; Feature selection; Multi-streams paradigm; CHMM; Dimensionality reduction; F-ratio

* Corresponding author.
E-mail address: w.abdulla@auckland.ac.nz (W.H. Abdulla).

0020-0255/$ - see front matter
doi:10.1016/S0020-0255(03)00162-2

1. Introduction

Speaker-independent speech recognition systems have many parameters to optimise during the implementation course. There are vast uncertainties to deal with, coming from varied production behaviour of different speakers. Statistical approaches using HMM show superiority over the other techniques to capture and model the features that are carrying the spoken information. The HMM framework interprets the speech signals to changeable-duration sequence of events called states [25]. The performance of the HMM model in discriminating the acoustic classes is highly affected by the observation feature vectors. They are considered as abstract mappings of the highly redundant speech samples. The feature vectors needed to be as short as possible in terms of dimensions which imply redundancy removal, and contain as much as possible of linguistic information. The selected features must assure fast training and recognition procedures, as well as superior acoustic class discrimination. The feature vectors have been widely investigated, and many suggestions have been proposed to reach the ultimate optimality goal of good abstraction and representation. The current approaches rely mainly on the successful Mel frequency cepstral coefficients (MFCCs) vectors to represent the speech samples. Other types of features, different from the MFCCs, have also been introduced and have some strength in certain applications. No feature set can be decided as the best absolute performer under all environmental conditions, in the automatic speech recognition (ASR) systems. One solution to exploit the strengths of the different feature sets is to combine them deliberately under a suitable paradigm. The combination of the features can be done at several points within the ASR structure. The features can be concatenated at the very beginning stage in the feature streams domain and presented to the general classifier, or left as they are in different independent streams and presented to separate classifiers. Also, the outcomes of the classifiers can be merged then presented to the general HMM decoder, or left as they are and presented to separate HMM decoders. The two main questions that needed to be answered in any multi-stream based ASR system design are: What feature set to stream? And, where to merge? There is no agreed analytical procedure to answer these queries, and they have mainly heuristic oriented solutions. However, there are some trends in using statistical notions to help in some decisions. The conditional mutual information (CMI) is used to predict which feature streams will merge most advantageously, and which of the many possible-merging strategies will be most successful, it answers the first question. The CMI of the raw feature streams is supposed to help in deciding whether to merge them together as one large stream, or to feed them separately into independent classifiers for later merging [7,8]. The results of the CMI technique are not very encouraging as reported.

The important property regarding the feature streams nature is that combining a number of diverse feature streams often improves the recognition performance, and the greatest benefits come from combinations between the most diverse features [30]. Different front-end structures can be used to maximise the feature stream diversity. Combining perceptual linear predictive (PLP) features with the modulation-filtered spectrogram (MSG) features improves the recognition rate significantly [9]. In fact, any change in the feature vectors preparation procedure will lead to an improvement in the recognition rate. Bella et al. [4] found that the combination of nearly identical sets of features with only difference in frame rate, which was set between 80 and 125 frames per second, was enough to introduce some decorrelation between the errors in the streams. The merged system performed significantly better than any one of the component streams. Variable frame rate is also useful in single stream ASR system. In this case, the frame rate is increased for rapidly changing segments with relatively high energy, and reduced for steady-state segments [31].

The classifiers are either Gaussian mixture models (GMM) or neural networks (NNs). Hybridising HMM with NN is widely used in single-stream structure for continuous speech recognition systems [22]. HMM speech recognition systems typically use GMM. The NNs are becoming popular in the multi-stream paradigm, due to their potential in estimating the probability functions and classifications. Merging the streams after the classification stage (posterior merging), rather than feature concatenation, ameliorates the recognition rate one step further [7,8,26]. The classifiers outcomes might be merged and decorrelated first, then presented to a GMM of a classical HMM decoder for better recognition performance [9]. The multi-estimation notion is also applicable to the NN based systems. It has been shown that recognition performance can be improved by using the same feature sets to train two NNs, with different initialisation points [19].

The other interesting approach in multi-streams research trends comes from sub-band notion. Rather than deriving the probability streams from completely different acoustic representations, it is also possible to divide a single representation into disjoint regions across the spectrum. Each of the sub-bands can then be used as the basis for separate probability estimators. The output of these estimators can be combined, either by averaging the log posterior probabilities for each class, or by using more complex methods including multi-layer perceptrons or weighted combinations [5,6]. More specifically, in the sub-band technique, the whole frequency band of the speech signal is split into several sub-bands. Then, each of these sub-bands is processed independently, mostly by a hybrid HMM neural network model. This technique is based on the assumption of sub-band independence, which is not very true, as in reality there is dependency between them. Next, the sub-band outcomes are recombined at several stages during the utterance period according to certain

criteria. The main advantage of this approach is the robustness of the recogniser to selective narrow-band noise [21]. This technique is also adopted in random field modelling to model the hidden states of HMM [12].

The multi-streaming is also viewed from another perspective by splitting the feature vectors into a specified number of sub-vectors, which are then processed by different quantizers, and a vector of discrete values with the same length as the number of sub-vectors is the input to the discrete recogniser [28]. This system was experimented with 9, 15, 24, and 39 sub-vectors, and it showed improvement in recognition rate as compared with the conventional CHMM.

The multi-stream approach was also investigated from the recognition rate in noisy environment perspective, and it showed substantial improvement in recognition performance under different noise sources [27].

In this paper, we will deal with the Mels coefficients and their first and second derivatives, as three independent streams. These streams have some sort of dependency, as it is obvious from the way of producing them. However, they showed enhanced effectiveness on the recognition rate when they were dealt with as independent. Thus, the feature vectors adopted comprise 39 coefficients (12 Mels and one power coefficient with their first and second derivatives) per observation; equally divided between three streams. Then we will reduce the dimensionality of each stream, using the F-ratio technique as a figure of merit. This reduction leaves only 28 MFCCs per observation vector to be used in our ASR system, instead of the original 39 coefficients.

This paper is organised as follows. Section 2 briefly demonstrates some related feature vectors designs. Section 3 describes the F-ratio as a figure of merit to assess the importance of the features, and how it can be directly applied on the HMM parameters. Section 4 explains the parallel HMM notion and the dimensionality-reduction application. Section 5 evaluates the performance of ASR systems based on different paradigms. The conclusions will be summarised in Section 6.

2. Feature vector design based on static and dynamic coefficients

The current approaches rely mainly on the successful Mel frequency cepstral coefficients (MFCCs) vectors to represent each 10–50 ms window of speech samples, taken each 5–25 ms, by a single vector of certain dimension. The window length and rate as well as the feature vectors dimension are decided according to the application task. For many applications the most effective components of the Mel scale features are the first 12 coefficients (excluding the zero coefficient), which are also called static coefficients. These coefficients are the features used by the HMMs to detect the stationary events in the speech signal spectra. Moreover, it has been shown that the speech recognition rate is noticeably improved when using additional coefficients, representing the

dynamic behaviour of the signal. These coefficients are the first and second derivatives of the cepstral coefficients of the static feature vectors [10,11,16,17]. The power coefficients, which represent the energy content of the signal and their first and second derivatives, have also important roles to be included in the representation of the feature vectors. The first and second derivatives are approximated by difference regression equations, and accordingly they are named delta and delta–delta coefficients or first and second deltas, respectively. The power coefficient, which represents the power of the signal within the processed windows, is concatenated with the Mel coefficients. The static coefficients are normally more effective in a clean environment, while the dynamic coefficients are more robust in a noisy environment [14]. Concatenating the static coefficients with their fist and second derivatives increases the recognition rate, but it has two drawbacks. First, the static coefficients will dominate the effect of the dynamic coefficients. Second, it increases the dimensionality of the feature vectors. Fig. 1 shows the power and Mel coefficients with their derivatives for the phoneme "O", and how the static coefficients dominance is apparent. This dominance lets the static coefficients be more effective than the first and second deltas during the measurement of the distances between the feature vectors, although the 1st and the second deltas might carry more information in certain parts of the signal. The distance measurement is the salient operation in all speech recognition algorithms. If we normalise the coefficients,

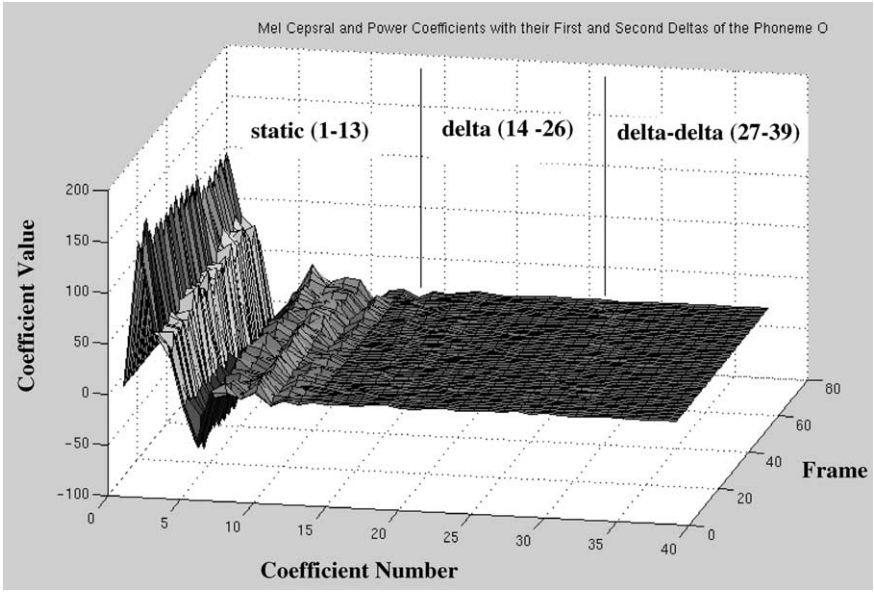

Fig. 1. The power and Mel-scale coefficients with their first and second derivatives.

as a remedy to this problem, we will misplace the actual weight of each coefficient within the feature vector.

One approach in the feature vectors design; a composite distance metric approach, was applied to accommodate for the relative importance and magnitude of different entities of the feature vectors [10,11,20]. The following distance metric "Dist" was used, in preparation of the Vector Quantization (VQ) bins for speech recognition system using HMM

$$\text{Dist} = \sum_{i=1}^{12}(C_i^{\text{r}} - C_i^{\text{t}})^2 + W_d \sum_{i=1}^{12}(D_i^{\text{r}} - D_i^{\text{t}})^2 + W_p(C_0^{\text{r}} - C_0^{\text{t}})^2 + W_{p'}(D_0^{\text{r}} - D_0^{\text{t}})^2 \tag{1}$$

where C_i represents an LPC cepstrum coefficient, D_i is a difference LPC cepstral coefficient (delta), C_0 is the power term and D_0 is the difference power. W_d, W_p, and $W_{p'}$ are empirically determined weighting factors that account for the relative importance and magnitude of the first difference coefficients (the second difference was not used). The superscripts 'r' and 't' refer to the reference and test vectors. The performance of the speech recogniser was improved according to this formula but the VQ distortion was still high.

Another more developed feature-vector design can be achieved by quantising each set of feature-vectors by a separate codebook which introduces multiple codebooks [15,17,18]. Multiple codebooks were introduced as a better option than the composite distance technique [13]. This technique outperformed the composite distance metric approach by reducing the quantization error, resulting from long feature vectors, which lead to better recognition rate. The multiple codebooks approach was adopted by the SPHINX speech recognition system based on the semi-continuous hidden Markov model (SCHMM) framework [18].

This technique belongs to a multi-streams paradigm, as the static and dynamic features are dealt with independently to find the observation probability distribution. The HMM should be modified in this case to produce multiple short observation vectors at each time instant, instead of single long observation vectors. The observations output probability should be modified by merging the probability of different streams to be suitable to embed in the HMM baseline, according to the formula

$$b_i(O_t) = \prod_c \sum_{k=1}^{L} p^c(O_t|v_k^c).b_i^c(v_k^c) \tag{2}$$

where $b_i(O_t)$ is the output probability of observation O_t given state i. $p^c(O_t|v_k^c)$ is the probability of observation O_t being in a codebook c and having a codeword v_k. $b_i^c(v_k^c)$ is the a priori probability of the codeword v_k of codebook c being in state i. L is the number of codewords in each codebook.

Our method tackles both the dominance and the dimensionality problems in a more effective way. It considers each stream as an independent feature vectors set and construct an HMM for each of them. The final responses of these models to any set of input vectors are combined to decide its class. Each feature vectors set is independently reduced in dimensionality, using the F-ratio technique. The advantage of our technique is justifiable by a statistical belief stated that combining multiple estimators for the same underlying value leads to better estimation.

3. Dimensionality reduction based on the F-ratio figures

The F-ratio is a measure that can be used to evaluate the effectiveness of a particular feature. It has been widely used as a figure of merit for feature selection in speaker recognition applications [24,29]. It is defined as the ratio of the between-class variance (B) and the within-class variance (W). In the contest of feature selection for pattern classification, the F-ratio can be considered as a strong catalyst to select the features that maximise the separation between different classes and minimise the scatter within these classes. The following assumptions have to be satisfied when using the F-ratio as a figure of merit for dimensionality reduction:

- The feature vectors within each class must have Gaussian distribution. This condition can be satisfied if we use sufficient training dataset, according to the central limit theorem.
- The features should be statistically uncorrelated. In practice this condition is hardly satisfied, and the correlated features can be transformed into uncorrelated features via suitable transformation such as the principal component analysis (PCA) and the linear discriminant analysis (LDA) techniques. However, if we use the Mel frequency cepstral coefficients to construct the feature vectors, then we can consider the feature vectors uncorrelated, since the discrete cosine transform (DCT) is used to prepare these vectors, which performs the adequate decorrelation.
- The variances within each class must be equal. Since the variances within each class are generally not equal, the pooled within-class variance is used to define the F-ratio.

The F-ratio technique can be formulated as follows [23]:

Let us consider that the number of training feature vectors, training patterns, in the jth class of K classes is N_j. Thus the F-ratio of the ith feature can be defined by

$$F_i = \frac{B_i}{W_i} \tag{3}$$

where B_i is the between-class variance and W_i is the pooled within-class variance of the ith feature, which can be mathematically defined by

$$B_i = \frac{1}{K}\sum_{j=1}^{K}(\mu_{ij} - \mu_i)^2 \tag{4}$$

$$W_i = \frac{1}{K}\sum_{j=1}^{K} W_{ij} \tag{5}$$

where μ_{ij} and W_{ij} are the mean and variance of the ith feature, respectively, for the jth class, and μ_i is the overall mean of the ith feature. These are given by

$$\mu_{ij} = \frac{1}{N_j}\sum_{n=1}^{N_j} x_{ijn} \tag{6}$$

$$W_{ij} = \frac{1}{N_j}\sum_{n=1}^{N_j}(x_{ijn} - \mu_{ij})^2 \tag{7}$$

$$\mu_i = \frac{1}{K}\sum_{j=1}^{K}\mu_{ij} \tag{8}$$

where x_{ijn} is the ith feature of the nth training feature vector, training pattern, from the jth class.

In our approach in using the F-ratio we make use of the HMM properties to facilitate the implementation of this technique in assessing and reducing the number of features. The HMM technique used is implicitly considering the Gaussian behaviour of the feature vectors which satisfies the first condition needed by the F-ratio method. The second condition is satisfied by using diagonal covariance within the structure of the HMM. Finally, the F-ratio averaging is conducted across all the models according to the formula

$$F^{\text{ave}} = \frac{1}{H}\sum_{i=1}^{H} F_i \tag{9}$$

where H represents the number of the HMMs.

The averaged F-ratio can be sorted into descending order, and then we can select the corresponding top Q features, which simply indicate the most prominent features within the whole set of features. Fig. 2 shows the mean F-ratio using this technique. If we sort the values of the resulting mean F-ratio in descending order, we can determine the features from the most prominent one to the least prominent one.

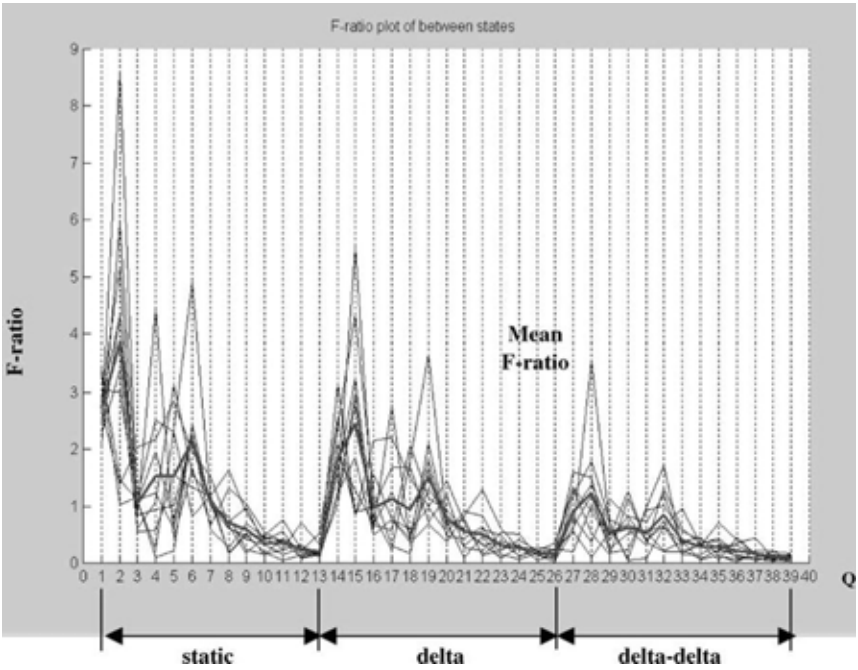

Fig. 2. F-ratio of several HMM models and their mean.

4. Parallel HMM multi-streams-based system

We developed this system based on the multi-stream notion, targeting the advantages of alleviating the dominance problem, the dimensionality reduction and flexibility of the design [3]. The speech signal feature vectors selected are the power and Mel-scale coefficients with their first and second derivatives, deltas. This selection is due to the high potential of these coefficients in carrying the static and temporal information of the spoken signals. The first derivative can be approximated by the regression formula

$$\Delta x(n) = \frac{\sum_{m=-N}^{N} mx(n+m)}{\sum_{m=-N}^{N} m^2} \tag{10}$$

where N is the delta period over which the difference is taken.

The second derivative, $\Delta\Delta$x, is approximated by applying the above equation on the resultant Δx(n). The spectral behaviours of the three components x, Δx, and $\Delta\Delta$x are different even they are derived from the same source, and the dependency is obvious from the way of derivation. This leads to the

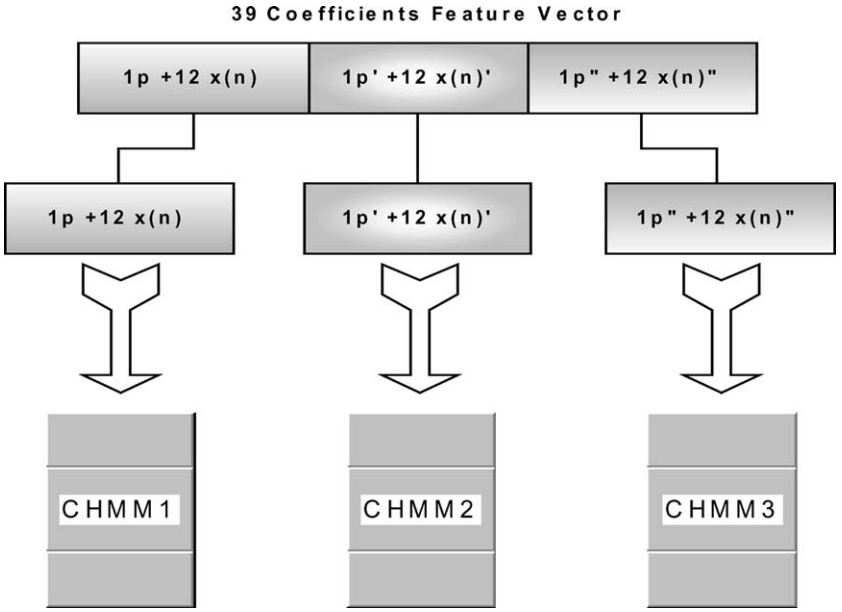

Fig. 3. Feature vector segmentation and processing. The prime and the double prime notations refer to the delta and delta–delta coefficients, respectively.

assumption that the stationary states of the three components are different from each other. To substantiate this assumption an experiment has been done, in which the three components of the feature vectors are considered to be independent, to model three CHMMs. The 39 coefficients feature vectors are extracted from the input speech signal: one power and 12 Mel coefficients with their first and second derivatives. Then this long vector is segmented into three 13-coefficient vectors to be dealt with as three independent components as shown in Fig. 3. Three streams CHMMs are trained on a certain word and the stationary states are backtracked for each stream using Viterbi algorithm [25].

Fig. 4 shows the states detected by the three CHMM models. This figure shows clearly that the states boundaries detected by the three models are not synchronised most of the time. This finding is due to the difference in the spectral characteristics of the states corresponding to each stream. There is no association between the states assigned to a certain feature in the different stream models. These states represent stationary classes of speech signal and there is also no correspondence between them and the linguistic units such as the phonemes and the sub-word classes.

The unsynchronised states are the reason for not adopting the technique of representing the output probability density by the sum of the individual log

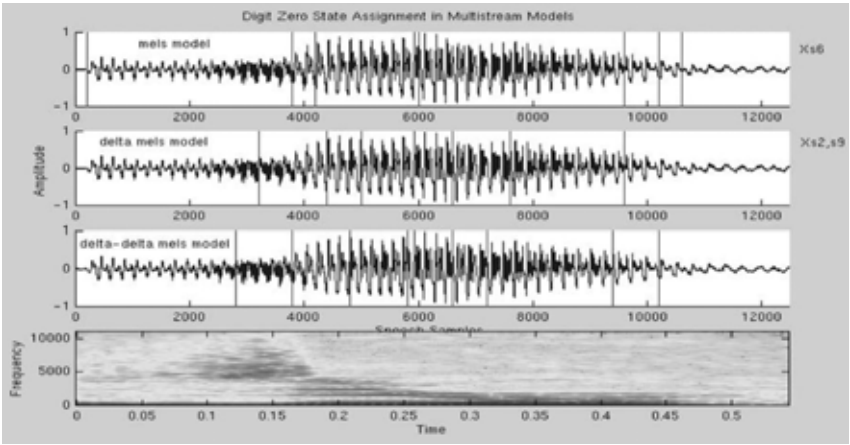

Fig. 4. State assignment by the three streams CHMM. The vertical lines are the states' boundaries. State 6 of Mels model and states 2 and 9 of the delta Mels model are skipped and indicated by Xs6, Xs2, and Xs9. This is due to the left-to-right CHMM structure, which allows one state skip.

probabilities of the streams at a certain time, as indicated in Eq. (2). A better choice would be to leave the streams behave as independent components (although, even this is not quite true), and merge the log likelihood at the end. The developed system has two phases to build: training and recognition. In the training phase, the topology used is left-to-right, allowing one state skip, and the type of the model is CHMM with nine states and five mixtures per state. Then three CHMMs have to be trained, one model per stream. During the recognition phase, the input speech signal is pre-processed and extracted from its background, first [1,2]. Then, each segment of the feature vectors is presented to its corresponding model. The output log-likelihood probabilities of the three streams are merged to decide the recognised word. The block diagram of the proposed system of the three streams is shown in
Fig. 5.

If the log-likelihood of each stream is represented by $P(O/\lambda^s)$, where s is 1, 2, or 3 for three streams, then the merging strategy can be achieved by two methods:

(a) The recognised word in this method is simply the weighted sum of the log probabilities of the three streams

$$P(O \mid system) = \sum_{s=1}^{3} w^s \; P(O \mid \lambda^s) \tag{11}$$

The weighting factor was taken to be equal in the developed system and also was adjusted heuristically according to the importance of the streams using the F-ratio technique, as will be discussed later in this section.

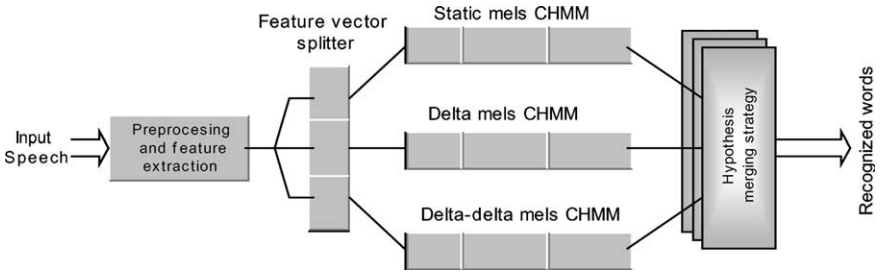

Fig. 5. Implemented multi-stream system structure.

(b) A more sophisticated technique using neural networks can be used. This technique takes into account the importance of each stream according to a certain maximisation criteria. Despite the superiority of this method over the previous one in part (a), we are not in favour of it. This is because it would need a larger training dataset, and it adds more parameters to the model. These parameters are increasing in the order of $O(mN)$, for one layer only, where N is the number of words to be recognised and m is the number of streams. The neural networks can be more beneficial to a system trained on the phonemes, as their number is limited, and it is expected to offer better optimisation decision than the simple heuristic weighted sum.

The structure depicted in Fig. 5 suggests investigating adding more streams from other features, like isolating the power coefficient with its deltas into independent stream.

The feature vectors dimension of the existing streams could be reduced by choosing the features that are more influential than the others, using the F-ratio technique discussed in Section 4. The F-ratio can also be used her as a figure of merit to identify the weight of each feature set in the classification property. Following the same method described in Section 4, we can plot the state F-ratio curves of the three streams and their means as in Fig. 6.

The features can also be sorted according to their classification importance to form Table 1. If we compare the F-ratio scores plotted in Figs. 2 and 6, we will see high coherence between them, which suggests the selection of the F-ratio as a successful figure of merit. In the top 28 ranked features the proportions of the Mel coefficients are $\{|Q_0| + |Q_1| + |Q_2| + |Q_3| = 3 + 10 + 9 + 6\}$, which is exactly the same result achieved from the single-stream experiments plotted in Fig. 2.

The F-ratio can also be used successfully in deciding the weight of each stream in the multi-stream paradigm systems. The average F-ratio of each stream has been taken and considered as the weight that decides the importance of this stream in the classification decision. Thus we will have the following weights corresponding to each stream:

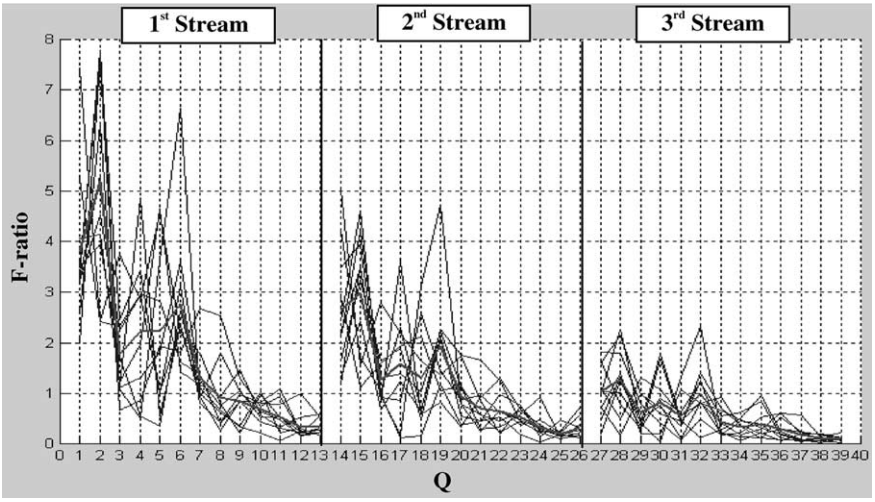

Fig. 6. State F-ratio of the multi-stream models. The thick red lines indicate the mean of the class F-ratio curves.

$$w^1 = 1.662 \quad \text{for the first stream (static)}$$

$$w^2 = 1.0531 \quad \text{for the second stream (delta)}$$

$$w^3 = 0.5549 \quad \text{for the third stream (delta–delta)}$$

We can also see from Fig. 6 that the mean state F-ratio scores for multi-stream case are higher than those indicated in Fig. 2 for single stream case. This indicates that the classification property of the multi-stream-based models outperforms its corresponding single stream counterpart.

5. Comparative studies of different ASR system paradigms

In this section, we depict the recognition rate of some successful ASR systems. The feature vectors used in all the systems are of 28 MFCCs and in the proportion specified in the previous section. The paradigms will be denoted by the following names for simplicity:

ASR-1: is a single stream multi-mixture model of type CHMM with nine states and five mixtures.

ASR-2: is a multi-stream ASR with equal stream merging weights. Three streams are used; each of them is modelled by a five mixtures CHMM with nine states. This system is shown in Fig. 5.

ASR-3: is the same as ASR-2 but uses the F-ratio merging weights.

Upon testing these systems under the same conditions and using three datasets we got the results indicated in Table 2. The datasets are collected from

Table 1
Multi-stream Mel frequency cepstral coefficients (MFFC) feature ordering using F-ratio as a figure of merit

Rank	Feature index	Corresponding coefficient	F-ratio value
1	2	C_1	5.1971
2	1	C_0	3.8122
3	15	ΔC_1	2.9792
4	6	C_5	2.7751
5	14	ΔC_0	2.6531
6	5	C_4	2.2297
7	4	C_3	2.2220
8	19	ΔC_5	1.9283
9	3	C_2	1.7663
10	17	ΔC_3	1.5716
11	18	ΔC_4	1.3029
12	7	C_6	1.3009
13	16	ΔC_2	1.2727
14	28	$\Delta\Delta C_1$	1.2622
15	27	$\Delta\Delta C_0$	1.0269
16	32	$\Delta\Delta C_5$	1.0012
17	20	ΔC_6	0.9454
18	8	C_7	0.9201
19	9	C_8	0.8401
20	30	$\Delta\Delta C_3$	0.8121
21	21	ΔC_7	0.6948
22	10	C_9	0.6748
23	22	ΔC_8	0.6519
24	29	$\Delta\Delta C_2$	0.6199
25	31	$\Delta\Delta C_4$	0.5449
26	23	ΔC_9	0.5314
27	11	C_{10}	0.5246
28	33	$\Delta\Delta C_6$	0.4313
29	35	$\Delta\Delta C_8$	0.3922
30	34	$\Delta\Delta C_7$	0.3351
31	13	C_{12}	0.3333
32	12	C_{11}	0.3312
33	24	ΔC_{10}	0.3266
34	26	ΔC_{12}	0.3134
35	36	$\Delta\Delta C_9$	0.2898
36	25	ΔC_{11}	0.2048
37	37	$\Delta\Delta C_{10}$	0.1782
38	38	$\Delta\Delta C_{11}$	0.1119
39	39	$\Delta\Delta C_{12}$	0.0958

the Otago Speech Corpus [1] and the number of words in each of them is indicated between brackets in the table.

[1] Otago Speech Corpus can be downloaded from http://kel.otago.ac.nz/hyspeech/corpus

Table 2
Recognition rates of four different ASR systems using three datasets

	Recognition rate (%)		
	DATASET-I (10)	DATASET-II (30)	DATASET-III (54)
ASR-1	100	97.3	92.5
ASR-2	100	99.2	97.3
ASR-3	100	99.2	98.9

Several conclusions can be deduced from Table 2. Perfect results obtained from all the ASR systems when tested by DATASET-I. Thus under this dataset we recommend ASR-1 system since it is the simplest paradigm. The multi-stream models, ASR-2 and ASR-3, show their superiority over the single stream one, ASR-1, when tested by DATASET-II and DATASET-III. The F-ratio merging weights based system, ASR-3, is the best among all the three systems when subjected to DATASET-III.

6. Conclusions

In this paper, we have investigated the problem of improving the speech recognition performance by restructuring the method of using the feature vectors. Instead of dealing with the composite static and dynamic speech signal features based on the MFCCs as a single stream, we proposed splitting them into three independent streams. Despite the fact that the streams' independence assumption is not very precise, from the method of deriving their vectors, the multi-streams paradigm has outperformed the baseline single-stream paradigm. This improvement is mostly predicated to two reasons. First, multi-streaming alleviates the dominance problem of any feature set over the others. Second, it reduces the dimensionality of the feature vectors used in each stream, which prevents the curse of the dimensionality problem. Merging the streams prematurely in an early stage, as discussed in Section 3, to construct one model only, would improve the performance of the recogniser over the single-stream based systems but not as much if we let each stream have its own model. This is due to the unsynchronised nature of the states among the streams, as depicted in Fig. 4, which makes the premature merging inefficient. Thus, in our paradigm we favour letting each stream's model detects its own stationary states. Then, the resultants of the parallel models are combined, under a certain hypothesis merging strategy. We demonstrated two merging strategies. One presumed equal-weight streams while the other suggested that the streams weights were proportional to the F-ratio average values. The latter had outperformed the first as depicted in Table 2. A NN can be introduced as an option in deciding the merging weights. However, we ruled out this option, because it would need bigger training data sets and it required more parameters

to optimise. The NN increases the number of parameters in the order of $O(mN)$, for one layer only, where N is the number of words to be recognised and m is the number of streams. The neural networks can be reconsidered on the phonemes based speech recognition systems, because the number of phonemes is limited and it is expected to offer better optimisation decision than the simple heuristic weighted sums.

The notion of the dimensionality reduction has been studied from another perspective. The F-ratio as a figure of merit in evaluating the feature importance was relied on, to select 28 MFFC features out of the full 39 features. In this case we have had 11, 10, and 7 features in the static and the two consecutive dynamic streams respectively. We applied the F-ratio technique directly on the HMM parameters rather than the usual long method based on the training data.

The potential of the multi-streams paradigm is very flexible and opens the door to investigate adding more streams from mechanisms other than MFCC features, such as the perceptual linear prediction coefficients (PLP). The increase of the dimensionality problem from the inclusion of extra features is not yet a problem in the multi-streams paradigm, as we can sensibly split the features and forward them to relevant streams.

However, the pitfall in this paradigm is in using the HMM decoding for as many streams as there are in the system. This problem can be diminished through using hardware processors to implement the HMM decoding. In this case, the whole processing time will be less than that of the single high-dimension stream because the streams' models are independent and working in parallel.

Acknowledgements

The authors would like to thank Professor Garry Tee for his constructive comments on this paper. This research is funded by The University of Auckland Research Fund under project number UARF 3602239/9273 and FRST of New Zealand, grant NERF AUT02/001.

References

[1] W.H. Abdulla, N.K. Kasabov, Speech recognition enhancement via robust CHMM speech background discrimination, in: Proc. ICONIP/ANZIIS/ANNES'99 International Workshop, New Zealand, 1999.

[2] W.H. Abdulla, N.K. Kasabov, Two pass hidden Markov model for speech recognition systems, in: Proceedings of ICICS'99, Singapore, 1999.

[3] W.H. Abdulla, N.K. Kasabov, Feature selection for parallel CHMM speech recognition systems, in: Proceedings of the Fifth Joint Conference on Information Science, vol. 2, 2000, pp. 874–878.

[4] J. Billa, T. Colhurst, et al., Recent experiments in large vocabulary conversational speech recognition, in: Proceedings of IEEE ICASSP'99, Phoenix, 1999.
[5] H. Bourlard, S. Bengio, et al., New approaches towards robust and adaptive speech recognition, in: T.K. Leen, T.G. Dietterich, V. Tresp (Eds.), Advances in Neural Information Processing Systems, vol. 13, MIT Press, Cambridge, MA, 2001, pp. 751–757.
[6] H. Bourlard, S. Dupont, et al., Multi-stream Speech Recognition, IDAP, 1996.
[7] D.P. Ellis, Improved recognition by combining different features and different systems, in: Proceedings of AVIOS-2000, San Jose, 2000.
[8] D.P. Ellis, Using mutual information to design feature combinations, in: Proceedings of the ICSLP-2000, Beijing, 2000.
[9] D.P. Ellis, R. Singh, et al., Tandem acoustic modelling in large-vocabulary recognition, in: Proceedings of the IEEE ICASSP'2001, Salt Lake City, 2001.
[10] S. Furui, Speaker independent isolated word recognition based on emphasized spectral dynamics, in: Proceedings of the IEEE ICASSP'86, Tokyo, Japan, 1986.
[11] S. Furui, Speaker independent isolated word recognition using dynamic features of speech recognition, IEEE Trans. ASSP 34 (2) (1986) 52–59.
[12] G. Gravier, M. Sigelle, et al., Markov random field modelling for speech recognition, Aust. J. Intell. Inform. Process. Systems 5 (4) (1998) 245–251.
[13] V.N. Gupta, M. Lenning, et al., Integration of acoustic information in a large vocabulary word recognizer, in: Proceedings of the IEEE ICASSP'87, 1987.
[14] B.A. Hanson, T.H. Applebaum, et al., Spectral dynamics for speech recognition under adverse conditions, in: C.-H. Lee, F.K. Soong, K.K. Paliwal (Eds.), Automatic Speech and Speaker Recognition, Kluwer Academic Publishers, Dordrecht, 1996.
[15] X.D. Huang, M.A. Ariki, et al., Hidden Markov Models for Speech Recognition, Edinburgh University Press, Edinburgh, 1990.
[16] X.D. Huang, K.-F. Lee, et al., Improved acoustic modeling for the SPHINX speech recognition system, in: Proceedings of the IEEE ICASSP'91, Toronto, Canada, 1991.
[17] M.Y. Hwang, Subphonetic acoustic modeling for speaker independent continuous speech recognition, CMU, 1993.
[18] M.Y. Hwang, X.D. Huang, Subphonetic modeling with Markov states-senone, in: Proceedings of the IEEE ICASSP'92, 1992.
[19] A. Janin, D.P. Ellis, et al., Multi-stream speech recognition: ready for prime time? in: Proceedings of the Eurospeech, Budapest, 1999.
[20] K.-F. Lee, Automatic Speech Recognition, Kluwer Academic Publishers, Dordrecht, 1989.
[21] J. Ming, F.J. Smith, A probabilistic union model for sub-band based robust speech recognition, in: Proceedings of the IEEE ICASSP'2000, Istanbul, Turkey, 2000.
[22] N. Morgan, H. Bourlard, Continuous speech recognition: an introduction to the hybrid HMM/connectionist approach, Signal Processing Magazine (May) (1995) 25–42.
[23] K.K. Paliwal, Dimensionality reduction of the enhanced feature set for the HMM-based speech recognizer, Digital Signal Processing 2 (1992) 157–173.
[24] S. Pruzansky, Talker recognition procedure based on analysis of variance, J. Acoust. Soc. Am. 36 (1964) 2041–2047.
[25] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.
[26] S. Sharma, D.P. Ellis, et al., Feature extraction using non-linear transformation for robust speech recognition on the Aurora database, in: Proc. of the IEEE ICASSP'2000, Istanbul, 2000.
[27] S. Tibrewala, H. Hermansky, Multi-stream approach in acoustic modelling, in: Proceedings of the LVCSR-Hub5 Workshop, 1997.
[28] S. Tsakalidis, V. Digalakis, et al., Efficient speech recognition using subvector quantization and discrete-mixture HMMs, in: Proceedings of the IEEE ICASSP'99, Phoenix, Arizona, 1999.

[29] J.J. Wolf, Efficient acoustic parameters for speaker recognition, J. Acoust. Soc. Am. 51 (1972) 2044–2056.
[30] S. Wu, B. Kingsbury, et al., Performance improvements through combining phone- and syllable-length information in automatic speech recognition, in: Proceedings of the ICSLP-98, Sydney, 1998.
[31] Q. Zhu, A. Alwan, On the use of variable frame rate analysis in speech recognition, in: Proceedings of the IEEE ICASSP'2000, Istanbul, Turkey, 2000.