IMPROVING SPEECH RECOGNITION PERFORMANCE BY USING MULTI-MODEL APPROACHES

Ji Ming, Philip Hanna, Darryl Stewart, Marie Owens, F. Jack Smith

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

ABSTRACT

Most current speech recognition systems are built upon a single type of model, e.g. an HMM or certain type of segment based model, and furthermore typically employs only one type of acoustic feature e.g. MFCCs and their variants. This entails that the system may not be robust should the modeling assumptions be violated. Recent research efforts have investigated the use of multi-scale/multi-band acoustic features for robust speech recognition. This paper described a multi-model approach as an alternative and complement to the multi-feature approaches. The multi-model approach seeks a combination of different types of acoustic model, thereby integrating the capabilities of each individual model for capturing discriminative information. An example system built upon the combination of the standard HMM technique with a segment-based modeling technique was implemented. Experiments for both isolated-word and continuous speech recognition have shown improved performances over each of the individual models considered in isolation.

1. INTRODUCTION

An acoustic model provides a mapping of acoustic-phonetic information. As such, the accuracy of the model has a fundamental effect on the performance and robustness of automatic speech recognition. Because different models make different simplifying assumptions, each specific model may only be capable of characterizing a certain aspect of the available information. Most current speech recognition systems are built upon a single type of model, e.g. an HMM or certain type of segment based model, and furthermore typically employ only one type of acoustic feature e.g. MFCCs and their variants. This entails that the system may not be robust should the modeling assumptions be violated. This is significant, as the production of more robust recognition systems is essential.

Ideally an acoustic model should be capable of capturing all of the discriminative information found in a given acoustic signal. Multi-feature and multi-model approaches offer a practical solution. The current research into the multi-feature techniques has investigated the calculation of multi-scale acoustic features in either the time or the frequency domain, and the combination of these feature streams within an HMM framework [1, 2, 8, 10]. Each feature stream represents a different characteristic of the input information. The combination of different feature streams has been accomplished by either directly creating an augmented feature vector that consists of all the component streams, or alternatively merging the likelihoods associated with each feature stream. Such systems have shown improved performance and robustness over the corresponding single feature stream based systems [1, 8, 10].

In this paper we investigate a multi-model approach as an alternative and complement to the multi-feature approach. The multi-model approach differs from the multi-feature approach in that it seeks a combination of different types of acoustic model, thereby integrating the capabilities of each individual model for capturing discriminative information. The proposed research is based on the observation that while the conventional HMM with multiple mixture densities is effective in representing the diversity of the static spectral characteristics, it is ineffective in capturing dynamic spectral information; likewise, while segment based models improve upon the standard HMM in terms of captured dynamic information, the inclusion of a segmental-level multiple mixture representation may prove detrimental due to the considerable increase in model complexity [9]. In other words, it may be assumed that there is no unique modeling method that encompasses the other methods in terms of the amount of information being captured. Should this assumption be true, then it is possible that an effective combination of different modeling techniques, with each technique emphasizing a different aspect of the input information, will result in a model that captures more information than any of the individual techniques considered in isolation. This research is significant in that it may bring about a significant improvement in the robustness of current speech recognition systems with relatively little effort.

2. A MULTI-MODEL APPROACH

For the creation of a multi-model system we need to address at least two issues: 1) which modeling techniques can be effectively combined, and 2) which methods can be used to effectively accomplish this combination. We focused our research on the possible combinations of HMM based techniques. Based on the above discussion, we suggest a combination of the standard HMM employing a multiple mixture of static densities with segment-based models, thereby integrating their capabilities for capturing both the static and dynamic spectral characteristics of speech. This combination is based on the assumption that there is little correlation between the error patterns that arise from each component model.

Given the component models, we investigated the use of an HMM framework to form the combined model. Specifically, we define the state-dependent observation densities of the combined model as the product of the corresponding densities from each of the component models, i.e.

$$b_i(x) = \prod_m b_i^m(x) \quad (1)$$

where $b_i^m(x)$ and $b_i(x)$ represent the observation densities of the *m*'th component model and the combined model respectively, for state *i*. If normalization of (1) is required then an exponential weighting can be introduced to each component density to balance their combination. Given (1), the likelihood function of the combined HMM can be written as

$$p(o|\lambda) = \sum_s \pi_{s_0} \prod_t a_{s_{t-1}s_t} \prod_m b_{s_t}^m(o_t) \quad (2)$$

where *o* is a time sequence of observations and λ is the parameter set of the combined model.

The model defined by (2) is equivalent to a linear combination of the component observation likelihood functions in the logarithmic domain, a method used by some multi-feature models for combining likelihoods from different feature streams (e.g. [1, 2, 8]). Of interest is the difference between (2) and those multi-feature methods. In (2) each $b_i^m(x)$ represents a different type of observation density and all the $b_i^m(x)$'s are applied to the same feature stream *o*; whilst in the multi-feature methods the same type of density is used for all the $b_i^m(x)$'s, with each $b_i^m(x)$ accounting for a different type of feature input.

The system structure shown in (2) has the advantage that it permits computationally effective training and decoding, and therefore retains one of the most attractive characteristics of the standard HMM technique. Following the standard procedure, a maximum-likelihood estimate of the model parameter set λ can be obtained by an iterative maximization of the following auxiliary function

$$Q(\lambda_0, \lambda) = \sum_s p(o, s|\lambda_0) \ln p(o, s|\lambda) \quad (3)$$

where λ_0 is an estimate from the previous iteration and $p(o, s|\lambda)$ is given by

$$p(o, s|\lambda) = \pi_{s_0} \prod_t a_{s_{t-1}s_t} \prod_m b_{s_t}^m(o_t) \quad (4)$$

Substituting (4) into (3) we obtain, particularly, an integral term of $Q(\lambda_0, \lambda)$ relating to the $b_i^m(x)$'s

$$Q(\lambda_0, \{b_i^m\}) = \sum_i \sum_t \sum_m p(o, s_t = i|\lambda_0) \ln b_i^m(o_t) \quad (5)$$

In (5), the probability $p(o, s_t = i|\lambda_0)$ for each *t* and *i* can be calculated using the standard forward-backward recursions. Hence, the re-estimation formula for the appropriate parameter vector of each $b_i^m(x)$, θ_i^m, is readily obtained by solving the corresponding equation

$$\sum_t p(o, s_t = i|\lambda_0) \frac{\partial b_i^m(o_t)}{\partial \theta_i^m} \cdot \frac{1}{b_i^m(o_t)} = 0 \quad (6)$$

These formulae permit the model parameters to be estimated in a computationally effective manner.

3. AN EXAMPLE SYSTEM

In this section we describe an implementation of the multi-model system (2) by using specific examples for the $b_i^m(x)$'s. We chose to combine the standard HMM employing a multiple mixture of Gaussian densities with a segment-based model, namely the inter-frame dependent HMM (IFDHMM) [4, 5, 7]. For the standard HMM, the *K*-mixture state-*i* observation density is given by

$$b_i^{std}(x) = \sum_{k=1}^{K} w_{ik} g_{ik}(x) \quad (7)$$

where $g_{ik}(x)$ is the *k*'th mixture component Gaussian and w_{ik} the corresponding mixture weight. The standard HMM with multiple mixture densities (7) is effective for representing the diversity of the static spectral characteristics of speech. However, it fails to adequately capture the dynamic spectral characteristics of speech, due to the frame independence assumption. During the past decade, various modified models have been proposed to overcome this problem [9]. Generally, a certain type of segment-level probability density is used to replace the initial frame-level density, thereby capturing longer-term dynamic spectral information. The IFDHMM embodies a modeling technique that we developed earlier as an alternative to the existing techniques for representing segmental level characteristics. The IFDHMM represents such characteristics by assuming that each acoustic frame is dependent upon a segment of preceding or succeeding frames. Specifically, the state-*i* observation density of the model is defined as [7]

$$b_i^{ifd}(x|x_1 \ldots x_N) = \sum_{n=1}^{N} c_{in} g_{in}(x|x_n) \quad (8)$$

where *N* defines the length of the conditional segment, $g_{in}(x|x_n)$ is a conditional Gaussian density capturing the correlation between *x* and the *n*'th conditional frame x_n, and c_{in} is the corresponding weight, satisfying the constraints $c_{in} \geq 0$ and $\Sigma_n c_{in} = 1$. The conditional Gaussian density function $g_{in}(x|x_n)$ can be shown to have a parametric form [7]

$$g_{in}(x|x_n) \propto \exp(-1/2 (x - H_{in} x_n - \mu_{in})' U_{in} (x - H_{in} x_n - \mu_{in})) \quad (9)$$

where μ_{in} is a *L*-dimensional vector and H_{in} and U_{in} are both $L \times L$ matrices, *L* being the dimensionality of the frame vector. Given an observation sequence *o*, the *N* conditional frames associated with each frame o_t, i.e. $o_{t-\tau(1)}, \ldots, o_{t-\tau(N)}$, are defined by a pre-chosen time-lag sequence $\tau(1), \ldots, \tau(N)$. Positive $\tau(n)$'s corresponds to a preceding-frame dependent system and negative $\tau(n)$'s corresponds to a succeeding-frame dependent system. Both models, along with the standard HMM (7), are combined according to (2) to form the combined model, i.e.

$$p(o|\lambda) = \sum_s \pi_{s_0} \prod_t a_{s_{t-1}s_t} \cdot b_{s_t}^{std}(o_t) \cdot b_{s_t}^{ifd}(o_t | o_{t-\tau(1)} \ldots o_{t-\tau(N)}) \cdot b_{s_t}^{ifd}(o_t | o_{t+\tau(1)} \ldots o_{t+\tau(N)}) \quad (10)$$

The combination of both the preceding and succeeding frame dependent models has been justified by our previous research in terms of improved performance [4, 5]. Given the non-stationary nature of speech, it is reasonable to assume that for a particular frame, the succeeding (or preceding) frames contain useful dynamic information that may not be encapsulated in the preceding (or succeeding) frames.

4. EXPERIMENTS

Experiments for both isolated-word and continuous speech recognition have been conducted. The isolated-word recognition experiments are based on a speaker-independent alphabetic database (provided by British Telecom Laboratories), from which the highly confusable E-set (b, c, d, e, g, p, t and v) is extracted. The database contains three repetitions of each word by a total of 104 speakers (52 male and 52 female). Among the 104 speakers, 52 were designated for training and the other 52 for testing. For each word, then, about 155 utterances are available for training, and a total of 1219 utterances are available for testing for all eight words. In addition to isolated word recognition, phone recognition experiments have been performed using the TIMIT database (1990 release). Following the recommendations by NIST [3], the database was subdivided into training and test sets, with the core test set being used in recognition. For the recognition of the E-set, a state-tied model topology using 15 states for each word, with the final 9 states tied among all the eight words, was adopted. For the phone recognition experiments, we built models for 48 phones and differentiated the standard 39-phone set [6]. Each phone was modeled with 3 states, a left-to-right topology and no context dependency. In all experiments, Mel-frequency cepstral coefficients (MFCCs) plus their first order differential parameters are calculated as the feature vector for each frame. Furthermore, all models used diagonal-type covariance matrices.

4.1 E-Set Recognition Results

The results presented in this section test the performance of the example system (10) for the recognition of the E-set. As described in Section 3, three component modeling techniques are combined in the system, namely a standard HMM and two IFDHMMs, one IFDHMM with a dependency upon preceding frames and the other with a dependency upon succeeding frames. As a starting point, Table 1 shows the recognition results of the individual component models. For the standard HMM, the results are presented as a function of the number of mixtures, and for the IFDHMMs the results are shown as a function of the number of conditional frames. For the IFDHMM, the number of conditional frames that is employed is directly proportional to the length of the segment being accounted for by the model. The results in Table 1 indicate that, due to an appropriate modeling of the longer-term dynamic spectra of speech, the IFDHMMs outperformed the standard HMM using multiple mixtures of static densities.

Next, we examine the performance of a simplified version of (10) by including only the two IFDHMM components in the model combination. The results are shown in Table 2, as a function of the number of conditional frames used in each component model.

Model	Parameter (*K* or *N*)	Accuracy (%)
Standard HMM	*K*=1	86.3
	K=3	88.8
	K=5	89.6
IFDHMM with preceding frame dependency	*N*=2	90.8
	N=3	91.7
	N=4	92.3
IFDHMM with succeeding frame dependency	*N*=2	90.8
	N=3	91.2
	N=4	91.6

Table 1. Recognition results of the individual models for the E-set. *K* and *N* represent the number of mixtures and the number of conditional frames used in each state in the appropriate model, respectively.

Model combination	Parameter (*N*) in each IFDHMM	Accuracy (%)
$ifd^- + ifd^+$	N=2	92.5
	N=3	93.0
	N=4	93.6

Table 2. Recognition results of a simplified combined model for the E-set. This model combines two IFDHMMs, one with a dependency upon preceding frames (ifd^-) and the other with a dependency upon succeeding frames (ifd^+). *N* is the number of conditional frames used in each component IFDHMM.

Comparing Table 2 with Table 1, it can be seen that the combined model always produces a higher accuracy than the corresponding component models operated individually. This phenomenon has already been reported previously [4, 5]. The non-stationary characteristics of speech entail that each of the two component IFDHMMs captures some useful dynamic spectral information that is not contained in the other. The combined model utilizes the information found in both component models. This led to the improved performance.

Finally, we include the standard HMM component into the model combination. The recognition results are shown in Table 3, where a fixed number of 4 conditional frames are used in each IFDHMM component, and the number of mixtures used in the standard HMM component is varied between 1 and 5. Comparing Table 3 with Table 1 and Table 2, we can see that the inclusion of a single-mixture, standard HMM component brought about little improvement in the performance. This is due to the poor accuracy of the single-mixture density in characterizing the static spectral variations. However, as the number of mixtures increased, the performance improvement due to the addition of the standard HMM component became significant. Typically, for the 4-conditional-frame and 5-mixture case, the error reduction resulting from the inclusion of the standard HMM component reached 24.7%, 25% and 17.2% for the (ifd^-+std), (ifd^++std) and (ifd^-+ifd^++std) model combinations respectively. Our best result, 94.7%, is obtained by the full implementation of the example system that combines all the three types of component model. Inevitably, compared to each individual model, the combined

model has an increased parameter size, but less so than a corresponding segmental-level multiple mixture model.

Model combination	Parameter (K) in standard HMM	Accuracy (%)
ifd^- + std	1	92.2
	3	93.9
	5	94.2
ifd^+ + std	1	92.3
	3	93.7
	5	93.7
ifd^- + ifd^+ + std	1	93.2
	3	94.0
	5	94.7

Table 3. Recognition results of the combined model for the E-set. The model combines the standard HMM (std) with IFDHMMs using preceding (ifd^-) and/or succeeding (ifd^+) frame dependencies. The number of conditional frames used for the IFDHMMs (N) is fixed at 4 and the number of mixtures used in the standard HMM (K) is varied as shown.

4.2 Phone Recognition Results

The results presented in this section test the performance of the example system (10) for the context-independent recognition of phones within the TIMIT database. In the experiments, the number of mixtures used for the standard HMM component was fixed at 16 and the number of conditional frames used for each of the IFDHMM components was fixed at 4. Additionally, a bigram phone language model was estimated on the training set and applied during recognition. Table 4 shows the phone recognition accuracies produced by the appropriate systems.

Model	Phone Recognition accuracy (%)
std	65.8
Ifd^- + ifd^+ + std	66.1

Table 4. Context-independent phone recognition results based on the TIMIT database.

While the combined model achieved significant error reductions in the recognition of the E-set, it obtained a less significant improvement in TIMIT phone recognition, as can be seen from Table 4. It is well known that phones have a short duration, cannot be pronounced in isolation, and that their characteristics greatly vary depending upon their context; thereby preventing an effective capture of sufficient and accurate dynamic spectral information. In order to obtain sufficient acoustic discrimination a larger unit than the phoneme (e.g. a syllable) is desirable.

5. SUMMARY

An acoustic model is a simplified mathematical representation of acoustic-phonetic information. The simplifying assumptions inherent to each model entail that it may only be capable of capturing a certain aspect of the available information. An effective combination of different types of model should therefore permit a combined model that can utilize all the information captured by the individual models. This paper presents some preliminary research in combining certain types of acoustic model for speech recognition. In particular, we designed and implemented a single HMM framework, which combines a segment-based modeling technique with the standard HMM technique. The experiments for both isolated-word and continuous speech recognition have shown that the combined model has the potential of producing a significantly higher performance than the individual models considered in isolation. The implemented model, though specific, may have a more general significance. That is, improved performance can be obtained by combining different types of acoustic model.

6. ACKNOWLEDGEMENT

This research is supported by the EPSRC under grant GR/K82505. Acknowledgement is also due to British Telecom Laboratories for providing the experimental database.

7. REFERENCES

[1] Bourlard, H., and Dupont, S. "A new ASR approach based on independent processing and recombination of partial frequency bands", *ICSLP'96*, Philadelphia, pp. 426-429, Oct. 1996.

[2] Dupont, S., and Bourlard, H. "Using multiple time scales in a multi-stream speech recognition system", *Eurospeech'97*, Rhodes, Greece, pp. 3-6, Sept. 1997.

[3] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*. NISTIR 4930, 1993.

[4] Hanna, P., Ming, J., O'Boyle, P., and Smith, F. J. "Modelling interframe dependence with preceding and succeeding Frames", *Eurospeech'97*, Rhodes, Greece, pp. 1167-1170.

[5] Hanna, P., Harte, N., Ming, J., Vaseghi, S., and Smith, F. J. "Variation of features of interframe dependent HMM for speech recognition", *IEE Electronics Letters*, Vol. 34, pp. 858-859, 1998.

[6] Lee, K.-F., and Hon, H.-W. "Speaker-independent phone recognition using hidden Markov models", *IEEE Trans. Acoust. Speech, Signal Processing*, Vol. 37, pp. 1641-1648, 1989.

[7] Ming, J., and Smith, F. J. "Modeling of the interframe dependence in an HMM using conditional Gaussian mixtures," *Computer Speech and Language*, Vol. 10, pp. 229-247, 1996.

[8] Okawa, S., Bocchieri, E., and Potamianos, A. "Multi-band speech recognition in noisy environments", *ICASSP'98*, Seattle, May 1998.

[9] Ostendorf, M., Digalakis, V. V., and Kimnall, O. A. "From HMMs to segment model: a united view of stochastic modeling for speech recognition", *IEEE Trans. Speech and Audio Processing*, Vol. 4, pp. 360-378, 1996.

[10] Tibrewala, S., and Hermansky, H. "Sub-band based recognition of noisy speech", *ICASSP'97*, pp. 1255-1258, Munich, Germany, May 1997.