CAPTURING DISCRIMINATIVE INFORMATION USING MULTIPLE MODELING TECHNIQUES

Ji Ming, Philip Hanna, Darryl Stewart, Saeed Vaseghi and F. Jack Smith

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

ABSTRACT

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 reports 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 recognition experiments, based on a speaker-independent E-set database, have shown that the combined model has the potential of producing a significantly higher performance than the individual models considered in isolation.

1. INTRODUCTION

Combining multiple information sources plays an important role in high performance speech recognition. In acoustic modeling, the extraction of different levels of discriminative information may be achieved in three ways: 1) different types of acoustic features, 2) different types of acoustic models, and 3) the combination of 1) and 2). The recent research efforts towards the first method have investigated the calculation and combination of multi-scale/multi-band acoustic features within an HMM [1, 2, 6, 8]. In these systems, 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 robustness over the traditional single feature stream based systems [1, 6, 8]. In this paper we investigate the second method, i.e. modeling an acoustic signal across multiple modeling techniques.

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 most current speech recognition systems are built upon a single modeling technique, e.g. an HMM or certain type of segment based model. While these techniques all aim to capture the useful discriminative spectral information contained in the speech utterance, they may each only capture a subset of the available information. For example, 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 [7]. 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 a union 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. In order to test the above hypothesis, we designed and implemented a single HMM framework, which combines a segment-based modeling technique with the standard HMM technique. 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. In addition, it is a useful complement to the current research in multiple-feature approaches, described above as method 1. Both approaches need to be advanced, and ultimately in the future they may be combined (i.e. method 3 above).

2. HMM BASED MULTI-MODEL TECHNIQUE

Currently the most successful ASR technique is based on HMMs and their variants. Therefore we focused our research on the creation of a single HMM framework, within which various HMM based techniques may be combined. The standard HMM technique has the advantages that it permits computationally effective algorithms for training and decoding, and additionally it offers a straightforward extension to multiple mixture densities, thereby considerably increasing the power of the model for representing the diversity of the static spectral characteristics of speech. However, the standard HMM fails to adequately model 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 [7]. 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. We suggest the 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.

2.1. A General Structure for the Combined Models

An HMM framework is employed to accomplish the above model combination. 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.

$$b_i(x) = \prod_m b_i^m(x) \tag{1}$$

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

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

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, 6]). 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. Both methods are common in that their effectiveness should depend on there being little correlation between the error patterns that arise from each component likelihood.

The model structure shown in (2) has the advantage that it permits computationally effective training and decoding, one of the most attractive characteristics of HMMs. In the following we show this by implementing (2) using specific examples for the $b_i^m(x)$'s.

2.2. A Specific Combined Model

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). The IFDHMM embodies a modeling technique that we developed earlier as an alternative to the existing techniques for representing segmental level characteristics [3-5]. For the standard HMM, the K-mixture state-i observation density is given by

$$b_i^{std}(x) = \sum_{k=1}^{K} w_{ik} g_{ik}(x) \tag{3}$$

where $g_{ik}(x)$ is the k'th mixture component Gaussian and w_{ik} the corresponding mixture weight. The IFDHMM represents the segment-level 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 [5]

$$b_i^{ifd}(x|x_1 ... x_N) = \sum_{n=1}^{N} c_{in} g_{in}(x|x_n) \tag{4}$$

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 [5]

$$g_{in}(x|x_n) \propto \exp(-1/2(x - H_{in}x_n - \mu_{in})' U_{in}(x - H_{in}x_n - \mu_{in})) \tag{5}$$

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. $o_{t-\tau(1)}$, …, $o_{t-\tau(N)}$, are defined by a pre-chosen time-lag sequence $\tau(1)$, …, $\tau(N)$. 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 (3), are combined according to (2) to form the combined model. The combination of both the preceding and succeeding frame dependent models has been justified by our previous research in terms of improved performance [3, 4]. 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.

Based on (2), we can write the likelihood function of the combined model as

$$p(o|\lambda) = \sum_s \pi_{s_0} \prod_t a_{s_{t-1}s_t} \cdot b_{s_t}^{std}(o_t) \cdot b_{s_t}^{ifd}(o_t | o_{t-\tau(1)} ... o_{t-\tau(N)}) \cdot b_{s_t}^{ifd}(o_t | o_{t+\tau(1)} ... o_{t+\tau(N)}) \tag{6}$$

Substituting (3) and (4) into (6), after some operator manipulation, it can be shown that

$$p(o|\lambda) = \sum_s \sum_\kappa \sum_\theta \sum_\nu p(o, s, \kappa, \theta, \nu | \lambda) \tag{7}$$

in which $p(o, s, \kappa, \theta, \nu | \lambda)$ is defined by

$$p(o,s,\kappa,\theta,\nu|\lambda) = \pi_{s_0} \prod_t a_{s_{t-1}s_t} \cdot w_{s_t k_t} g_{s_t k_t}(o_t) \cdot c_{s_t n_t} g_{s_t n_t}(o_t | o_{t-\tau(n_t)}) \cdot c_{s_t m_t} g_{s_t m_t}(o_t | o_{t+\tau(m_t)}) \tag{8}$$

and κ, θ and ν represent the *T*-tuples $(k_1,...,k_T)$, $(n_1,...,n_T)$ and $(m_1,...,m_T)$, respectively, with every k_t defined over (1, …, *K*), and every n_t and m_t defined over (1, …, *N*), assuming that the same number of conditional frames are used to model the preceding and succeeding frame dependencies. Based on (7), a maximum-likelihood estimate of the model parameter set λ can be obtained by an iterative maximization of the following auxiliary function

$$Q(\lambda_0,\lambda) = \sum_{s,\kappa,\theta,\nu} p(o,s,\kappa,\theta,\nu|\lambda_0) \ln p(o,s,\kappa,\theta,\nu|\lambda) \tag{9}$$

where λ_0 is an estimate from the previous iteration. This maximization can be accomplished using the standard forward-backward procedure, leading to the computationally effective model re-estimation algorithm.

3. EXPERIMENTS

The 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 for the experiments. The database contains three repetitions of each word by a total of 104 speakers; the database is roughly balanced with respect to age and gender. 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. The speech, sampled at 20 kHz, was divided into 25.6 ms frames with a consecutive frame overlap of 15 ms. Each frame is passed through a filter bank of 27 band-pass mel-frequency filters, from which 12 MFCCs plus their first order differential parameters are extracted. A state-tied model topology, using 15 states for each word and with the final 9 states tied among all the eight words, was used throughout the experiments. Furthermore, all models used diagonal-type covariance matrices.

3.1. Performance of the Individual Models

The results presented in this section test the performance of the individual models. As described above, three component modeling techniques are considered, namely a standard HMM and two IFDHMMs, one IFDHMM with a dependency upon preceding frames and the other with a dependency upon succeeding frames.

Table 1 shows the recognition results. 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. For the IFDHMM, the number of conditional frames being employed is directly proportional to the length of the segments being accounted for by the model. The results shown in Table 1 provide good evidence that an appropriate modeling of the longer-term dynamic spectra of speech is at least as important as the representation accuracy of the instantaneous spectra, achieved through the use of multiple mixtures of static densities.

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 performances of the standard HMM and IFDHMMs. The results are shown as a function of the number of mixtures (K) or the number of conditional frames (N) used in each state in the appropriate model.

3.2. Performance of the Combined Models

The combination of the above models has been tested using the algorithms described in Section 2. Firstly, we examine the effectiveness of the combination of the two IFDHMMs, one model employing preceding frame dependencies and the other succeeding frame dependencies. The results are shown in Table 2, as a function of the number of conditional frames used in each component model. Compared to 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 [3, 4]. 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.

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

Table 2: Recognition performance of the model combining two IFDHMMs, one with a dependency upon preceding frames (ifd^-) and the other with a dependency upon succeeding frames (ifd^+). The results are shown as a function of the number of conditional frames (N) used in each component model.

Next, 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 observe 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 $(\mathrm{ifd}^-$+std), $(\mathrm{ifd}^+$+std) and $(\mathrm{ifd}^- + \mathrm{ifd}^+$+std) model combinations respectively. Inevitably, compared to each individual model, the above combined models have an increased parameter size, but less so than a corresponding segmental-level multiple mixture model.

As the database used in this paper has also been used by many other researchers, a comparison between our results and others is made possible. To the authors' knowledge the previous highest accuracy was that of 94.6%, reported by Valtchev based on maximum mutual information estimation, applied to an HMM using full covariance matrices [9]. We achieved a similar result (94.7%) with our new combined model using a less complicated training algorithm.

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 performance of the model combining 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. Conclusions

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 which could be used 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 based on the combined model have shown an significantly improved performance over each of 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.

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.

REFERENCES

1. Bourlard, H., and Dupont, S. "A new ASR approach based on independent processing and recombination of partial frequency bands", ICSLP'96, pp. 426-429.
2. Dupont, S., and Bourlard, H. "Using multiple time scales in a multi-stream speech recognition system", Eurospeech'97, pp. 3-6.
3. Hanna, P., Ming, J., O'Boyle, P., and Smith, F. J. "Modelling interframe dependence with preceding and succeeding Frames", Eurospeech'97, pp. 1167-1170.
4. 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.
5. 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.
6. Okawa, S., Bocchieri, E., and Potamianos, A. "Multi-band speech recognition in noisy environments", ICASSP'98.
7. 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. SAP,* Vol. 4, pp. 360-378, 1996.
8. Tibrewala, S., and Hermansky, H. "Sub-band based recognition of noisy speech", ICASSP'97, pp. 1255-1258, 1997.
9. Valtchev, V. *Discriminative methods in HMM-based speech recognition,* PhD Dissertation, Cambridge University, England, 1995.