IMPROVED PHONE RECOGNITION USING BAYESIAN TRIPHONE MODELS

Ji Ming and F. Jack Smith

School of Electrical Engineering and Computer Science
The Queens University of Belfast, Belfast BT7 1NN, UK

ABSTRACT

A crucial issue in triphone based continuous speech recognition is the large number of models to be estimated against the limited availability of training data. This problem can be relieved by composing a triphone model from less context-dependent models. This paper introduces a new statistical framework, derived from the Bayesian principle, to perform such a composition. The potential power of this new framework is explored, both algorithmically and experimentally, by an implementation with hidden Markov modeling techniques. This implementation is applied to the recognition of the 39-phone set on the TIMIT database. The new model achieves 74.4% and 75.6% accuracy, respectively, on the core and complete test sets.

1. INTRODUCTION

Building triphonic models for continuous speech recognition has not been an easy task due to the data sparsity problem. Previous studies have attacked this problem by using model-interpolation [6] and quasi-triphone [7] techniques. In the model-interpolation technique, an under-trained triphone is re-tuned by interpolating the model with others of less context-dependency, i.e. the left and right bi-phone and the mono-phone models, which can be trained more reliably. This technique can improve the robustness of the models and the weights, balancing the combination, have been determined either by hand-tuning [9] or by using the deleted-interpolation algorithm [6]. The quasi-triphone model is based on a left-to-right HMM structure and on an assumption that the contexts mainly affect the outer states of an HMM. Therefore the first and last states are trained to distinguish the left and right contexts, respectively, and the central states can be assumed to be context-independent. This technique typically reduces the number of distinct models to be estimated from $\sim O(N^3)$ to $\sim 2O(N^2)$, where N is the number of phones. In addition, various context-clustering techniques for sharing training data from similar context-effects have been proposed. In HMM based systems, parametric tying has been excised from states [11] to mixture components [4] and to feature parameters [10]. Context clustering has been combined into both the model-interpolation and quasi-triphone systems to improve the models' trainability [6][7].

Building triphones based on model-interpolation involves heuristics and/or intensive computation in determining the interpolation weights. Besides, it can be argued that because this method separately estimates the component models and their interpolation weights, it results in typically sub-optimal models. The quasi-triphone model, on the other hand, is inaccurate for some short phones such as stops, affricates and some fricatives, which often have time durations no longer than a single frame of the normal length. In other words, the left and right context-effects are potentially temporally inseparable.

In this paper we introduce a new statistical framework for constructing triphone models from models of less context-dependency. This composition reduces the number of distinct models by higher than an order of magnitude and is therefore of great significance in triphone-based continuous speech recognition. The new approach is suggested with a hope to overcome the above mentioned problems. It is distinguished from the previous models in that it is built on the Bayesian principle, rather than on a heuristic method. The potential power of the new model is demonstrated by an implementation based on HMM techniques.

2. THEORETICAL FRAMEWORK

Denote by x is a phone-level observation and (a^-, a, a^+) a triphone unit, with a being some phone and a^- and a^+ being its left and right contexts, respectively. The problem of triphonic acoustic modeling can be expressed as the estimation of the probability density function (pdf) $p(x \mid a^-, a, a^+)$, of x generated from (a^-, a, a^+). Using the Bayesian principle

$$p(x \mid a^-, a, a^+) = \frac{p(a^-, a^+ \mid a, x)p(a, x)}{p(a^-, a^+ \mid a)p(a)} \qquad (1)$$

If we assume that: 1) a^- and a^+ are independent given a, i.e. $p(a^-, a^+ \mid a) = p(a^- \mid a)p(a^+ \mid a)$, and 2) a^- and a^+ are independent given a and x, i.e. $p(a^-, a^+ \mid a, x) = p(a^- \mid a, x)p(a^+ \mid a, x)$, (1) becomes

$$p(x \mid a^-, a, a^+) = \frac{p(a^- \mid a, x)p(a^+ \mid x, a)p(x, a)}{p(a^- \mid a)p(a^+ \mid a)p(a)} \qquad (2)$$

Therefore, by multiplying both the numerator and denominator of (2) by $p(x, a)p(a)$ it follows that

$$p(x \mid a^-, a, a^+) = \frac{p(x \mid a^-, a)p(x \mid a, a^+)}{p(x \mid a)} \qquad (3)$$

(3) indicates a novel way of obtaining a triphone model by composing models of less context-dependency, i.e. $p(x \mid$

This work was supported by EPSRC Grant GR/K82505

$a^-, a)$, $p(x \mid a, a^+)$ and $p(x \mid a)$, which correspond to the pdfs of x given the left-context, right-context and context-independent units, respectively. This composition leads to a reduction of the number of models to be estimated from $\sim O(N^3)$ to $\sim 2O(N^2)$. The assumptions made above in obtaining (3) simply mean that all combinations of the left and right contexts are permitted in forming the triphones. Without an appropriate language constraint, this turns out to be an inherent characteristic of all the approaches producing a triphone from the combination of left and right biphones [7]. Because the derivation of (3) is closely related to Bayesian statistics, we call (3) the Bayesian triphone model.

3. HMM BASED IMPLEMENTATION

3.1. The acoustic model

The Bayesian triphone model has been implemented by defining each of the component models, i.e. $p(x \mid a^-, a)$, $p(x \mid a, a^+)$ and $p(x \mid a)$ in (3), as an HMM. Let $x = (x_1, \ldots, x_T)$ denote a phone-level observation sequence and $-$, $+$ and $\sim$ differentiate the left-context, right-context and context-independent models, respectively. The likelihood function associated with the Bayesian triphone model can be written as

$$p(x \mid \lambda) = \frac{p(x \mid \lambda^-)p(x \mid \lambda^+)}{p(x \mid \lambda^\sim)} \tag{4}$$

where $\lambda = (\lambda^-, \lambda^+, \lambda^\sim)$ is the triphone model parameter set and each $p(x \mid \lambda^c)$, $c = -$, $+$ and $\sim$, is defined by

$$p(x \mid \lambda^c) = \sum_s \pi^c_{s_0} \prod_{t=1}^{T} a^c_{s_{t-1}s_t} b^c_{s_t}(x_t) \tag{5}$$

(5) is the standard HMM representation where λ^c is the model parameter set and $s = (s_0, \ldots, s_T)$ is the state sequence. (4) can be simplified by tying the state sequences among the left-context, right-context and context-independent models. The tying of the state-sequences is defined as $(\pi^-, A^-) = (\pi^+, A^+) = (\pi^\sim, A^\sim) = (\pi, A)$, and $p(s \mid x, \lambda^-) = p(s \mid x, \lambda^+) = p(s \mid x, \lambda^\sim) = p(s \mid x, \{\pi, A\})$, where π and A are the tied initial-state and state-transition probabilities, respectively. In other words, we assume that the state sequence of a given signal is dependent only on the nature of the signal; different models accounting for the same observation signal generate identical state-sequences. Substitute (5) into (4) and apply the tying of the state-sequences as defined above, it can be shown that (4) can be reduced to

$$p(x \mid \lambda) = \sum_s \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \frac{b^-_{s_t}(x_t) b^+_{s_t}(x_t)}{b^\sim_{s_t}(x_t)} \tag{6}$$

(6) is the triphone model which we implemented for phone recognition.

Specifically, assume that each $b^c_i(x)$ ($c = -$, $+$ and $\sim$) in (6) is a mixture Gaussian density of a form

$$b^c_i(x) = \sum_n w^c_{in} b^c_{in}(x) \tag{7}$$

where $b^c_{in}(x)$ is the nth Gaussian component in state i and w^c_{in} the corresponding weight. Substitute (7) into (6), note that $1/\sum_n w^\sim_{s_t n} b^\sim_{s_t n}(x_t) = \sum_n w^\sim_{s_t n} b^\sim_{s_t n}(x_t)/b^\sim_{s_t}(x_t)^2$ and that $\prod_{t=1}^{T} \sum_n w^c_{s_t n} b^c_{s_t n}(x_t) = \sum_{n_1 \ldots n_T} \prod_{t=1}^{T} w^c_{s_t n_t} b^c_{s_t n_t}(x_t)$, we therefore can write $p(x \mid \lambda)$ as

$$p(x \mid \lambda) = \sum_s \sum_{\mathcal{N}} \sum_{\mathcal{M}} \sum_{\mathcal{K}} p(x, s, \mathcal{N}, \mathcal{M}, \mathcal{K} \mid \lambda) \tag{8}$$

where $p(x, s, \mathcal{N}, \mathcal{M}, \mathcal{K} \mid \lambda)$ is defined as

$$p(x, s, \mathcal{N}, \mathcal{M}, \mathcal{K} \mid \lambda) = \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \cdot w^-_{s_t n_t} b^-_{s_t n_t}(x_t) w^+_{s_t m_t} b^+_{s_t m_t}(x_t) \frac{w^\sim_{s_t k_t} b^\sim_{s_t k_t}(x_t)}{b^\sim_{s_t}(x_t)^2} \tag{9}$$

and $\mathcal{N}$, $\mathcal{M}$ and $\mathcal{K}$ represent the T-tuples $(n_1, \ldots, n_T)$, $(m_1, \ldots, m_T)$ and $(k_1, \ldots, k_T)$, respectively, whose summations are over all possible $(n_1, \ldots, n_T)$s, $(m_1, \ldots, m_T)$s and $(k_1, \ldots, k_T)$s, respectively.

3.2. The forward-backward re-estimation algorithm

Following the usual practice, a maximum-likelihood estimate of λ, based on the likelihood function $p(x \mid \lambda)$ defined in (8), can be achieved by an iterative maximization of a Baum's auxiliary function

$$Q(\lambda, \hat{\lambda}) = \sum_{s, \mathcal{N}, \mathcal{M}, \mathcal{K}} p(x, s, \mathcal{N}, \mathcal{M}, \mathcal{K} \mid \lambda) \ln p(x, s, \mathcal{N}, \mathcal{M}, \mathcal{K} \mid \hat{\lambda}) \tag{10}$$

with respect to $\hat{\lambda}$ for a given previous estimate λ. Maximizing $Q(\lambda, \hat{\lambda})$ against parameters of the left and right context components is straightforward, resulting in the respective re-estimation formula. For example, a new estimate of the mean vector of b^-_{in}, occurring as a critical point of $Q(\lambda, \hat{\lambda})$, is given by

$$\hat{m}^-_{in} = \frac{\sum_{t=1}^{T} \xi^-_{in}(t) \cdot x_t}{\sum_{t=1}^{T} \xi^-_{in}(t)} \tag{11}$$

where $\xi^-_{in}(t)$ is short for the probability $p(x, s_t = i, n_t = n \mid \lambda)$, calculated by

$$\xi^-_{in}(t) = \sum_j \alpha_{t-1}(j) a_{ji} \frac{w^-_{in} b^-_{in}(x_t) b^+_i(x_t)}{b^\sim_i(x_t)} \beta_t(i) \tag{12}$$

where $\alpha_t(i)$ and $\beta_t(i)$ are the forward and backward probabilities, respectively, computed with the following recursions

$$\alpha_t(j) = \sum_i \alpha_{t-1}(i) a_{ij} \cdot \frac{b^-_j(x_t) b^+_j(x_t)}{b^\sim_j(x_t)} \tag{13}$$

and

$$\beta_t(i) = \sum_j \beta_{t+1}(j) a_{ij} \frac{b^-_j(x_{t+1}) b^+_j(x_{t+1})}{b^\sim_j(x_{t+1})} \tag{14}$$

(11)-(12) apply to b^+_{im} by interchanging the indexes between n and m and $-$ and $+$.

The re-estimation formula for the context-independent component is obtained by maximizing (10) against $\{\hat{b}^\sim_{ik}\}$.

Specifically, the partial derivative of $Q(\lambda, \hat{\lambda})$ against the mean vector of $\hat{b}^{\sim}_{ik}$ can be shown as

$$\frac{\partial Q(\lambda,\hat{\lambda})}{\partial \hat{m}^{\sim}_{ik}} = \sum_{t=1}^{T} \left[\xi^{\sim}_{ik}(t) - 2\xi_i(t) \frac{\hat{w}^{\sim}_{ik}\hat{b}^{\sim}_{ik}(x_t)}{\hat{b}^{\sim}_{i}(x_t)} \right] \frac{1}{\hat{b}^{\sim}_{ik}(x_t)} \frac{\partial \hat{b}^{\sim}_{ik}(x_t)}{\partial \hat{m}^{\sim}_{ik}} \tag{15}$$

where $\xi^{\sim}_{ik}(t)$ and $\xi_i(t)$ are short for the probabilities $p(x, s_t = i, k_t = k \mid \lambda)$ and $p(x, s_t = i \mid \lambda)$, respectively, and they are calculated by

$$\xi^{\sim}_{ik}(t) = \sum_j \alpha_{t-1}(j) a_{ji} \frac{b^{-}_i(x_t) b^{+}_i(x_t) w^{\sim}_{ik} b^{\sim}_{ik}(x_t)}{b^{\sim}_i(x_t)^2} \beta_t(i) \tag{16}$$

and

$$\xi_i(t) = \sum_j \alpha_{t-1}(j) a_{ji} \frac{b^{-}_i(x_t) b^{+}_i(x_t)}{b^{\sim}_i(x_t)} \beta_t(i) \tag{17}$$

An approximation to Equation (15) can be made by representing $\hat{w}^{\sim}_{ik}\hat{b}^{\sim}_{ik}(x_t)/\hat{b}^{\sim}_{i}(x_t)$ within the bracket in terms of the previous estimates. As such, note from (16) and (17) that $\xi_i(t) w^{\sim}_{ik} b^{\sim}_{ik}(x_t)/b^{\sim}_i(x_t) = \xi^{\sim}_{ik}(t)$, we therefore can write (15) as

$$\frac{\partial Q(\lambda,\hat{\lambda})}{\partial \hat{m}^{\sim}_{ik}} = -\sum_{t=1}^{T} \frac{\xi^{\sim}_{ik}(t)}{\hat{b}^{\sim}_{ik}(x_t)} \frac{\partial \hat{b}^{\sim}_{ik}(x_t)}{\partial \hat{m}^{\sim}_{ik}} \tag{18}$$

(18), as being set to zero and solved for $\hat{m}^{\sim}_{ik}$, resulting in the re-estimation equation

$$\hat{m}^{\sim}_{ik} = \frac{\sum_{t=1}^{T} \xi^{\sim}_{ik}(t) \cdot x_t}{\sum_{t=1}^{T} \xi^{\sim}_{ik}(t)} \tag{19}$$

In a similar way, we can obtain the re-estimation formula for the mixture weights and covariance matrices.

As can be seen, the above algorithm constructs the three component models and their composition in one step, subject to a common optimality criterion. This constitutes a distinguishing characteristic for the new model, as compared with the model-interpolation based approaches. Specifically, this characteristic makes the new model computationally efficient and, presumably, globally optimal.

3.3. Parametric tying with the new model

The problem of tying parameters within the new model is raised to improve the trainability of the model's biphone components. In particular, two strategies of state-level tying have been studied as a complement to the above training algorithms. Firstly, a tied-mixture structure [4] is introduced to the corresponding states of all the three component models accounting for the triphones of a phone. In such a model, state codebooks containing mixture densities are shared across the component models, while each component model has a distinct mixture weight distribution, which is specific to the respective context phone and the context independency. The re-estimation algorithm described above can be easily modified to accommodate this model. Specifically, a new estimate of the weight in each component model can be shown as

$$\hat{w}^{c}_{in} = \frac{\sum_{t=1}^{T} \xi^{c}_{in}}{\sum_{n'} \sum_{t=1}^{T} \xi^{c}_{in'}} \qquad c = -,\ +,\ \sim \tag{20}$$

where ξ^{c}_{in}s are defined in (12) and (16) respectively, with each $b^{c}_{in}(x) = b_{in}(x)$, and a new estimate of the mean vector of $b_{in}(x)$ is given by

$$\hat{m}_{in} = \frac{\sum_{t=1}^{T} \left(\xi^{-}_{in}(t) + \xi^{+}_{in}(t) - \xi^{\sim}_{in}(t) \right) \cdot x_t}{\sum_{t=1}^{T} \left(\xi^{-}_{in}(t) + \xi^{+}_{in}(t) - \xi^{\sim}_{in}(t) \right)} \tag{21}$$

Next, merging the context-specific weight distributions within the left and right biphones of a phone is introduced to the above tied-mixture model. This merging spans the same states of the models and accounts for those biphone weights trained with too few occurrences. The merging is based on the increase in the weighted-by-counts entropy [6]. To retain the context resolution, a threshold, indicating the minimum number of training samples needed to estimate a weight distribution, is introduced to stop the merging. Merging the weight distributions is performed following an iterative sorting-merging method. At each step of the iteration, the least frequent weight-distribution is merged upwards into another distribution chosen to minimize the entropy increase A new weight-distribution is then formed by combining the counts of occurrences of the merged distributions. The new distribution set is sorted again by frequency for the next step of merging. This sorting-merging process is repeated until the frequency of the least frequent distribution reaches the pre-defined threshold. An advantage of this sorting-merging method is that the weight-distributions will not be merged if they each have already satisfied the threshold, therefore retaining reasonable model resolution.

4. EXPERIMENTS

Experiments are performed with the TIMIT database (1990 release, [3]). Following convention, we recognize the standard 39-phone set. The database is subdivided into training and test sets based on the recommendations by NIST. Both the core and complete test sets are used in the experiments.

The Bayesian triphone model with tied-mixture states is implemented, with the merging of the context-specific mixture-component weights as an option. A simple HMM structure, with 3 states and a left-to-right topology, is used throughout the modeling. The codebook size for each tied state is chosen to be 16, each codeword being a Gaussian density with a diagonal covariance matrix. The speech signal is divided into frames, each with a length of 20 ms and adjacent frames overlapped by 10 ms. Ten Mel-frequency cepstral coefficients (MFCCs) and one normalized logarithmic energy, along with their first and second order differential versions defined over a window of $\pm$20 ms, are calculated as the observation vector for each frame. The models are initialized by first training a context-independent HMM for each phone. Afterwards, each required left and right context model is initialized by cloning the corresponding context-independent model. These serve as the initial component models for composing the Bayesian triphone models. Then, for each training sentence, the embedded training of the Bayesian triphone models is performed using the algorithm described in Section 3. Three embedded training iterations are run in each experiment. A bigram phone language model is estimated on the training set and is applied to the recognition experiments.

Table I and Table II show the recognition results of the new triphone model on the core and complete test sets, respectively. These results are produced by the models with and without merging the context-specific mixture weights. For merging the mixture weights, two thresholds, 50 and 100, are used, respectively, each setting a bottom number of training samples required to estimate a mixture-weight distribution. Since the TIMIT training set contains a significant number of both left and right biphones with very low frequency of occurrences, many weight distributions will be under-trained. This lack of robustness can be improved by an appropriate merging of the similar weight distributions, leading to an improvement in the recognition performance. This is seen in both Table I and Table II.

The comparison between our results and some of the best results reported previously by other researchers is summarized in Table III. The comparison is made on the same test set whenever the corresponding results are available, and for comparison, the results produced by the context-independent HMMs (mono-phones) are also included. To the authors' knowledge, the accuracies of 74.4% and 75.6%, obtained by the new model on the core and complete test sets respectively, are higher than those so far reported in the literature.

5. CONCLUSIONS

This paper introduced a new statistical framework for constructing triphonic models from models of less context-dependency. This composition reduces the number of models to be estimated by higher than an order of magnitude and is therefore of great significance in relieving the dtat sparsity problem in triphone-based continuous speech recognition. The potential power of the new framework has been explored by an implementation with the HMM technique. It is shown that the new model structure leads to efficient model estimation and optimization. Phone recognition experiments on the TIMIT database have shown improved accuracy over that obtained by other systems.

6. REFERENCES

[1] Chen, R. and Jamieson, L. H. "Explicit modeling of coarticulation in a statistical speech recognizer", ICASSP'96, pp. 463-466.

[2] Deng, L. and Sameti, H. "Transitional speech units and their representation by regressive Markov states: application to speech recognition", IEEE Trans. Speech and Audio Processing, vol. 4, pp. 301-306, 1996.

[3] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D. and Dahlgren, N. DARPA TIMIT Acoustic-phonetic continuous speech corpus CD-ROM. NISTIR 4930, 1993.

[4] Huang, X. "Phoneme classification using semicontinuous hidden Markov models", IEEE ASSP, vol. 40, pp. 1062-1067, 1992.

[5] Lamel, L. and Gauvain, J. "High performance speaker-independent phone recognition using CDHMM", EUROSPEECH'93, pp. 121-124.

[6] Lee, K. F. "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition", IEEE Trans. ASSP, vol. 38, pp. 599-609, 1990.

[7] Ljolje, A. "High accuracy phone recognition using context clustering and quasi-triphonic models", Computer Speech and Language, vol. 8, pp. 129-151, 1994.

[8] Robinson, A. "An application of recurrent nets to phone probability estimation", IEEE Trans. Neural Networks, vol. 5, pp. 298-305, 1994.

[9] Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M. and Makhoul, J. "Context-dependent modeling for acoustic-phonetic recognition of continuous speech", ICASSP'85, pp. 1205-1208.

[10] Takami, J. and Sagayama, S. "Four-level tied-structure for efficient representation of acoustic modeling", ICASSP'95, pp. 520-523.

[11] Young, S. and Woodland, P. "State clustering in HMM-based continuous speech recognition", Computer Speech and Language, vol. 8, pp. 369-384, 1994.

Table I. Phone recognition performance (%) of the new triphone model on the core test set

Merging threshold	Corr.	Acc.	Sub.	Del.	Ins.
No merging	76.8	72.9	17.3	5.9	3.9
50	77.7	74.0	16.3	6.0	3.7
100	77.8	74.4	16.0	6.2	3.4

Table II. Phone recognition performance (%) of the new triphone model on the complete test set

Merging threshold	Corr.	Acc.	Sub.	Del.	Ins.
No merging	78.6	74.9	15.5	5.9	3.7
50	79.0	75.6	15.1	5.9	3.4
100	79.0	75.6	15.0	6.0	3.4

Table III. Comparison of phone accuracy (%) between the new model and some previous models

New model	Some previous models	
Accuracy	Model	Accuracy
$74.4^{\dagger}$, $75.6^{\ddagger}$	Quasi-triphone [1]	70.4
	Gender-specific [5]	$71.1^{\dagger}$, $73.4^{\ddagger}$
	State clustering [11]	72.3
	Polynominal state [2]	$73.5^{\dagger}$
	Recurrent neural net [8]	$73.9^{\dagger}$, $75.0^{\ddagger}$
Mono-phone	$64.9^{\dagger}$, $66.0^{\ddagger}$	

† Core test set result.

‡ Complete test set result.