A BAYESIAN TRIPHONE MODEL WITH PARAMETER TYING

Ji Ming, Marie Owens, and F. Jack Smith *
Department of Computer Science, The Queens University of Belfast
Belfast BT7 1NN, Northern Ireland, UK
e-mail: j.ming, m.owens, fj.smith@qub.ac.uk

ABSTRACT

This paper introduces a new statistical framework for constructing triphonic models from models of less context-dependency. The new framework is derived from Bayesian statistics, and represents an alternative to other triphone-by-composition techniques, particularly to the model-interpolation and quasi-triphone approaches. The potential power of this new framework is explored by an implementation based on the hidden Markov modeling technique. It is shown that the new model structure includes the quasi-triphone model as a special case, and leads to more efficient parameter estimation than the model-interpolation method. Two strategies of state-level tying have been investigated within the new model structure. Phone recognition experiments on the TIMIT database show an increase in the accuracy over that obtained by other systems.

1 INTRODUCTION

A key issue in triphone based continuous speech recognition is the large number of parameters to be estimated against the limited availability of training data. In previous years, various approaches have been proposed to attack this problem. These approaches typically include model-interpolation, quasi-triphone and various parametric tying strategies. In the model-interpolation method [1], an under-trained triphone is re-tuned by interpolating the model with others of less context-dependency, i.e. the left-context, right-context and/or context-independent models, which can be trained more reliably. This technique can improve the robustness of the models and the interpolation weights for balancing the combination have been determined either by hand-tuning or by using deleted interpolation [1]. The quasi-triphone model [2] 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. We refer to these two methods as *triphone-by-composition* since both approximate triphonic context dependency via a composition of less context dependency. In addition, various *parametric tying* methods have been proposed in HMM based systems. These techniques, including various types of state tying (e.g. [3]) and mixture tying (e.g. [4]), approximate triphonic context dependency by sharing training data from similar context-effects.

Given limited availability of training data, the parametric tying methods may lose significant context resolution due to extensive clustering of the parameters to meet the acoustic robustness. This loss of context resolution can be considerably lower in the triphone-by-composition methods because only biphonic context dependency is to be estimated and therefore a smaller degree of tying is needed. However, a drawback of the triphone-by-composition method on its own is that it may reduce the accuracy of the triphone model when there are sufficient data to estimate a triphone accurately. A better solution, thus, would be a combination of the parametric tying and triphone-by-composition methods by using a generalized backoff mechanism. Rather than backing off a triphone directly to a biphone, as is implied in many parametric tying systems to account for a shortage of training, the generalized mechanism backs off a triphone to a composed triphone, therefore retaining a reasonable context sensitivity. The obvious difficulty is how to formulate the composition. This will be the focus of this paper.

Constructing triphone models based on model-interpolation involves heuristics and/or intensive computation in determining the interpolation weights. 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 composing a triphone model from models of less context-dependency. The new model is

*This work was supported by EPSRC Grant GR/K82505

suggested as an alternative to the above methods hoping to overcome the above mentioned problems. It is distinguished from the previous models in that it is built on Bayesian statistics, rather than on a heuristic method.

2 THE BAYESIAN TRIPHONE MODEL

Assuming that x is a phone-level acoustic 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 rule

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

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

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

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

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

(3) indicates a novel way of approximating a triphone model by composing models of less context-dependency, i.e. $p(x \mid a^-, a)$, $p(x \mid a, a^+)$ and $p(x \mid a)$, which correspond to the pdf's of x given the left-context-dependent (LCD), right-context-dependent (RCD) and context-independent (CI) units, respectively. This composition leads to a reduction of the number of models to be estimated from $\sim O(N^3)$ to $\sim 2O(N^2)$, without loss of context coverage. The assumptions made above in obtaining (3) simply mean that all combinations of the left and right contexts are permitted in forming the triphones. This causes no problem for training but it does cause a problem for recognition, as in other triphone-by-composition methods, by producing some illegal triphones. The effect of these extra triphones can be limited by trigram phonotactic constraints, as addressed in [2]. Because the derivation of (3) is based on Bayesian statistics, we call (3) the Bayesian triphone model.

(3) is more a framework than a specific model. In other words, it can be applied to many existing acoustic modeling techniques by associating the component pdf's with the respective model-based likelihood function. In the following we describe the implementation of this model by using standard hidden Markov modeling techniques.

3 ACOUSTIC MODELLING

Each component pdf on the right-hand side of (3) is associated with a corresponding acoustic model, which we assume to be an HMM. Let $x = (x_1, \ldots, x_T)$ denote a phone-level observation sequence and $-$, $+$ and $\sim$ differentiate the LCD, RCD and CI models, respectively, we have for each model the state-based likelihood function

$$p(x \mid s^c, \lambda^c) = \prod_{t=1}^{T} b^c_{s^c_t}(x_t) \tag{4}$$

where $c = -$, $+$ and $\sim$. (4) is the standard HMM representation where λ^c is the HMM parameter set and $s^c = (s^c_0, \ldots, s^c_T)$ is the state sequence. Given (4), the corresponding likelihood function associated with the Bayesian triphone model can be expressed as

$$\begin{aligned} p(x \mid s^-, s^+, s^\sim, \lambda) &= \frac{p(x \mid s^-, \lambda^-)p(x \mid s^+, \lambda^+)}{p(x \mid s^\sim, \lambda^\sim)} \\ &= \prod_{t=1}^{T} \frac{b^-_{s^-_t}(x_t) b^+_{s^+_t}(x_t)}{b^\sim_{s^\sim_t}(x_t)} \end{aligned} \tag{5}$$

where $\lambda = (\lambda^-, \lambda^+, \lambda^\sim)$ is the triphone model parameter set. Without loss of generality, we assume that the three component models, accounting for the same observation x, generate identical state-sequences subject to a common initial-state probability vector $[\pi_i]$ and state-transition probability matrix $[a_{ij}]$. So (5) is reduced to

$$p(x \mid s, \lambda) = \prod_{t=1}^{T} \frac{b^-_{s_t}(x_t) b^+_{s_t}(x_t)}{b^\sim_{s_t}(x_t)} \tag{6}$$

where s represents the common state-sequence shared by the three component models. Subsequently, the state-averaged likelihood function of the model is given by

$$p(x \mid \lambda) = \sum_s \pi_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \frac{b^-_{s_t}(x_t) b^+_{s_t}(x_t)}{b^\sim_{s_t}(x_t)} \tag{7}$$

(7) is the triphone model which we implemented in this paper. It is noted that this model includes the quasi-triphone model as a special case. Typically, considering a 3-state HMM for each component model. Using the above notation, the LCD, RCD and CI state-based probabilities are described by $[b^-_1, b^-_2, b^-_3]$, $[b^+_1, b^+_2, b^+_3]$ and $[b^\sim_1, b^\sim_2, b^\sim_3]$, respectively. The composed probabilities, as indicated in (7), are given by

$$\frac{b^-_1 b^+_1}{b^\sim_1}, \quad \frac{b^-_2 b^+_2}{b^\sim_2}, \quad \frac{b^-_3 b^+_3}{b^\sim_3}$$

In the above, if we assume that the last two states of the LCD model are context independent, i.e. $b^-_2 = b^\sim_2$ and $b^-_3 = b^\sim_3$, and that the first two states of the RCD model are context independent, i.e. $b^+_1 = b^\sim_1$ and $b^+_2 = b^\sim_2$, we then end up with the composed state-based probabilities

by the new model as $[b_1^-, b_2^\sim, b_3^+]$. This turns out to be the state probability topology assumed in the quasi-triphone model.

Assume that each $b_i^c(x)$ ($c = -$, $+$ and $\sim$) in (7) is a mixture Gaussian density of a form

$$b_i^c(x) = \sum_n w_{in}^c b_{in}^c(x) \tag{8}$$

where $b_{in}^c(x)$ is the nth Gaussian component in state i and w_{in}^c the corresponding weight. Substitute (8) into (7), note that $1/\sum_n w_{s_t n}^\sim b_{s_t n}^\sim(x_t) = \sum_n w_{s_t n}^\sim b_{s_t n}^\sim(x_t)/b_{s_t}^\sim(x_t)^2$ and $\prod_{t=1}^T \sum_n w_{s_t n}^c b_{s_t n}^c(x_t) = \sum_{n_1,\ldots,n_T} \prod_{t=1}^T w_{s_t n_t}^c b_{s_t n_t}^c(x_t)$, we therefore can write $p(x \mid \lambda)$ as

$$p(x \mid \lambda) = \sum_s \sum_{\mathcal{N}} \sum_{\mathcal{M}} \sum_{\mathcal{K}} p(x, s, \mathcal{N}, \mathcal{M}, \mathcal{K} \mid \lambda) \tag{9}$$

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

$$p(x, s, \mathcal{N}, \mathcal{M}, \mathcal{K} \mid \lambda) = \pi_{s_0} \prod_{t=1}^T a_{s_{t-1} s_t} \cdot w_{s_t n_t}^- b_{s_t n_t}^-(x_t) w_{s_t m_t}^+ b_{s_t m_t}^+(x_t) \frac{w_{s_t k_t}^\sim b_{s_t k_t}^\sim(x_t)}{b_{s_t}^\sim(x_t)^2} \tag{10}$$

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. The summations for $\mathcal{N}$, $\mathcal{M}$ and $\mathcal{K}$ are over all possible $(n_1, \ldots, n_T)$s, $(m_1, \ldots, m_T)$s and $(k_1, \ldots, k_T)$s, respectively. A forward-backward re-estimation algorithm can be developed for estimating the model defined above. Following the standard practice, a maximum-likelihood estimate of λ, based on the likelihood function $p(x \mid \lambda)$ defined in (9), can be achieved by an iterative maximization of a *Baum's auxiliary function*

$$Q(\lambda, \hat{\lambda}) = \sum_{s, \mathcal{N}, \mathcal{M}, \mathcal{K}} p(x, s, \mathcal{N}, \mathcal{M}, \mathcal{K} \mid \lambda) \ln p(x, s, \mathcal{N}, \mathcal{M}, \mathcal{K} \mid \hat{\lambda}) \tag{11}$$

with respect to $\hat{\lambda}$ for a given previous estimate λ. Maximizing $Q(\lambda, \hat{\lambda})$ against parameters of the LCD, RCD and CI components results in their respective re-estimation formula (a more detailed description of the algorithm may be found in [9]). The above algorithm constructs the LCD, RCD and CI component models and their composition in one step. This constitutes a potential advantage of the new model structure in terms of computational efficiency, as compared with the model-interpolation based approaches. The traditional interpolation model structure constructs a state-i observation probability using a form $b_i = \sum_n \lambda_{in} b_{in}$, where b_{in} is the n'th component probability and λ_{in} the interpolation weight. The interpolation weights are estimated separately from the component probabilities using, for example, deleted-interpolation on deleted blocks of training data [1].

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 algorithm. In the first strategy, 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, the state-i observation density of each component model can be expressed as

$$b_i^c(x) = \sum_n w_{in}^c b_{in}(x) \qquad c = -,\ +,\ \sim \tag{12}$$

where the $b_{in}(x)$s are the state-dependent mixture-component densities (state codewords), shared across all the component models covering the triphones of a phone; w_{in}^-s, w_{in}^+s and $w_{in}^\sim$s are the left-context, right-context and context-independent mixture-component weights, specific to their respective context phone and the context independency. 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 accounts for those biphone weights trained with too few occurrences. The merging is based on the increase in the weighted-by-counts entropy [1] and is stopped by a threshold indicating the minimum number of training samples required to estimate a weight distribution.

4 EXPERIMENTS

Experiments are performed with the TIMIT database. Following convention, we recognize the standard 39-phone set. 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 CI HMM for each phone. Afterwards, each required LCD and RCD model is initialized by cloning the corresponding CI 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 Bayesian 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. All the models being compared to are based on continuous densities and are context dependent. The comparison is made on the same test set whenever the corresponding results are available. 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

A new statistical framework for constructing triphonic models from models of less context-dependency is introduced. 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 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.

References

[1] K-F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE ASSP*, vol. 38, pp. 599-609, 1989.

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

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

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

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

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

[7] L. Deng and H. Sameti, "Transitional speech units and their representation by regressive Markov states: application to speech recognition," *IEEE SAP*, vol. 4, pp. 301-306, 1996.

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

[9] J. Ming and F. J. Smith, "Improved phone recognition using Bayesian triphone models," *ICASSP-98.*

Table I. Phone recognition results (%) 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 results (%) 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 other context-dependent models for recognizing TIMIT 39-phone set

	Test set		
Model	Core	Complete	Other
Quasi-triphone [5]			70.4
Gender-specific [6]	71.1	73.4	
State clustering [3]			72.3
Polynominal state [7]	73.5		
Recurrent neural net [8]	73.9	75.0	
New Bayesian triphone	74.4	75.6	