Improved Bayesian Training for Context-Dependent Modeling in Continuous Persian Speech Recognition

S.M. Ahadi

Electrical Engineering Department,
Amirkabir University of Technology, Tehran, Iran.
sma@aut.ac.ir

Abstract:

Context-dependent modeling is a widely used technique for better phone modeling in continuous speech recognition. While different types of context-dependent models have been used, triphones have been known as the most effective ones. In this paper, a Maximum *a Posteriori* (MAP) estimation approach has been used to estimate the parameters of the untied triphone model set used in data-driven clustering. The use of better prior parameters derived from two sets of more reliably trained biphone models has helped in this process. The result is better parameter tying where the tied-state triphone system built in this manner outperforms a similar system in which ordinary Maximum Likelihood (ML) approach was used to estimate the untied triphone system parameters. The technique may also be useful in other tying schemes used in context-dependent modeling.

Keywords: Bayesian training, Prior parameter estimation, Context-dependent modeling, Triphones, Biphones, State tying, Continuous speech recognition

1. Introduction

It is a widely known issue that co-articulation, i.e. the effect of neighboring sounds in the pronunciation of a phone, can radically change its acoustic properties. The speech recognition systems that rely on phoneme-based models may hence be adversely affected by this phenomenon. In order to alleviate this problem, context-dependent modeling has been used as the choice in continuous speech recognition. Many different types of context-dependent models have so far been implemented, including biphones, triphones, quinphones, etc. [1] [2]. Furthermore, other sub-word models that can partly model context, such as syllables and demi-syllables, have also been used [3] [4] [5]. However, triphones have gained much attention as one of the best performing context-dependent modeling methods.

While context-dependent models are more accurate models in comparison to context-independent ones, an important issue dealing with these models is found to be their large number in a continuous speech recognition system. Furthermore, to obtain a more accurate recognition system, one needs to use relatively complex output distributions, such as mixture-Gaussians, which can give a further substantial increase to the number of parameters within the system.

Several efforts have aimed to overcome the main problem of context-dependent and mixture-Gaussian modeling, i.e. the large number of system parameters that can result in poor parameter estimation. Most of the available triphone systems utilize a tying approach to reduce the overall number of system parameters and obtain robust parameter estimations [6] [7]. Others have used techniques such as model-interpolation and quasi-triphones to try help train better triphone models [8] [9]. These techniques have shown certain improvements in overall triphone system performance with tying approaches leading the list. It is also worth mentioning that in recent years, the issue of further improving the context-dependent models has not been dealt with extensively as the tying approaches are believed to have taken the context-dependent system performance close to its limits. The approaches introduced in more recent researches have also generally tried to improve the performance of tied-parameter systems.

Tying, itself, involves different approaches. Several levels of tying have been introduced in order to reduce the overall number of parameters in the system [10]. Starting from model level, lower levels of tying have been more successful due to their better sharing of details among different parts of speech units. In a HMM-based recognition system these include states, state transition probabilities, mixture components and even pdf parameters such as mean vectors, covariance matrices and mixture weights. Taking into account that lower level tying involves more complication in programming, tying in state and mixture component levels have been used most widely in speech recognition systems.

Among more recent efforts to improve the context-dependent systems performance, in [11] and [12] the syllabic structure of Persian and Thai languages have been explored to improve context dependent modeling. Willet *et al.* have used concept analysis as a mathematical means for improving tree-based state clustering [13]. López de Ipiña *et al.* used a data massaging feature in decision tree clustering to emphasize the data and a fast and efficient Growing and Pruning algorithm for the decision tree construction [14].

An initial step in the process of tying is clustering. The so-called data driven approach tries to provide an initial set of triphone parameters using available training data [10]. Then, clustering is performed based on this set of parameters. Obviously, due to the lack of training data, this approach suffers from weak clustering. An alternative approach uses a tree clustering approach based on the phonetic properties of the target language [2]. This approach has the advantage of not depending only on the initial inappropriately trained models, but it also needs a carefully designed clustering tree for the target language. Bayesian approaches such as MAP estimation, due to their higher performance in sparse data conditions, have also been used for parameter estimation in similar situations. It has been shown that starting from monophones as priors in a MAP estimation framework, better triphone models can be obtained in comparison to maximum likelihood estimation [15] [16]. Another Bayesian approach has been introduced that uses a common optimality criterion to construct triphone models using models of less context-dependency, i.e. left and right biphones and monophones [17]. Due to smaller number of parameters in such models, they have more robustly-trained parameters and hence can lead to better triphone modeling.

While parameter tying is a well known and vastly used approach in context-dependent modeling, in a data-driven tying approach a set of initial triphones are needed in order to enable us perform clustering. This is usually obtained by first cloning monophones, i.e. using the parameters from the same monophone model for all the models of its allophones. Then, in a subsequent training phase, these triphone models are trained using the available training data. Approaches such as monophone model cloning and inhibiting parameter updates during training, in the cases of the availability of very few observations, lead to more improved triphones. The resultant models are then used as the basic set for clustering and tying [10]. The major drawback of this approach is that due to the large number of parameters in the triphone system and in spite of utilizing such approaches as monophone cloning and update inhibition for few examples, the training of the initial triphone models may not be robust, which can adversely affect the clustering process. This has been the main reason for using MAP estimation in place of ML for this initial training stage, which is also believed to have a considerable impact on the tying process.

In this paper, we have shown that using enhanced prior models for the MAP estimation approach can even lead to better initial triphone models and consequently better tied models. The prior model set is obtained using an approach similar to that followed to create a quasi-triphone model set [9]. Although backing off to simpler context dependent models, i.e. monophones and biphones, has been introduced before for the construction of more detailed models, such as triphones [18], we have introduced a specific backing off approach to construct improved prior parameters for the MAP estimation process. Due to the desirable characteristics of MAP estimation that can start from prior models and gradually improve as more training data appears, the triphones trained in this manner are expected to perform better in comparison to ML-trained triphones. This can, in turn, lead to a better data-driven tied-parameter triphone system.

2. MAP versus ML Estimation

The Bayesian framework for the estimation of HMM parameters has this advantage over Maximum Likelihood (ML) estimation that makes use of a set of prior parameters. Given a set of observation sequences, $\boldsymbol{O} = \{\boldsymbol{o}_1,\ldots,\boldsymbol{o}_T\}$, the maximum likelihood estimate for the HMM parameter set (say λ) is found by setting

$$\frac{\partial}{\partial\lambda} P(\boldsymbol{o}_1,\ldots,\boldsymbol{o}_T \mid \lambda) = 0. \qquad (1)$$

In a Bayesian framework, however, the Maximum *a posteriori* (MAP) estimate for the HMM parameter set is found by maximizing the posterior distribution of parameters given the set of observation sequences ($\boldsymbol{O}$), i.e.

$$\frac{\partial}{\partial\lambda} P(\lambda \mid \boldsymbol{o}_1,\ldots,\boldsymbol{o}_T) = 0. \qquad (2)$$

According to the well-known Bayes rule,

$$P(\lambda \mid \boldsymbol{O}) = \frac{P(\boldsymbol{O} \mid \lambda)P(\lambda)}{P(\boldsymbol{O})}, \qquad (3)$$

where $P(\lambda)$ is a prior distribution function for the set of model parameters. Using the above concept, it has been shown that the MAP estimation for the HMM parameters can be found using a set of prior parameters [19]. As an example, the MAP estimation for mean parameters of the mixture-Gaussian pdfs is given by

$$\tilde{\boldsymbol{m}}_{ik} = \frac{\tau_{ik}\boldsymbol{\mu}_{ik} + \sum_{t=1}^{T}\gamma_{ikt}\boldsymbol{o}_t}{\tau_{ik} + \sum_{t=1}^{T}\gamma_{ikt}}, \qquad (4)$$

where $\boldsymbol{\mu}_{ik}$ is the prior mean parameter, τ_{ik} is a weighting prior parameter and γ_{ikt} is the probability of being in state i and mixture component k at time t given the observation $\boldsymbol{o}_t$.

Equation (4) can be directly compared to the mean estimation equation within a ML framework, such as Baum-Welch estimation. This comparison would result in that the first parts of both numerator and denominator of Equation (4) are the main difference between these two approaches, as far as the pdf mean parameters are concerned. As mentioned, these are the parts related to the prior parameters involved in MAP estimation. This also means that the MAP estimate for the mean parameter can be interpreted as a weighted interpolation between the ML estimate and the prior value of the mean. If both the prior mean value ($\boldsymbol{\mu}_{ik}$) and the prior weight (τ_{ik}) are chosen appropriately, then with small amounts of γ_{ikt},

which is in fact the mixture component (or state, in a single-Gaussian system) occupation count, the prior mean dominates. With increasing the available training data and the value of γ_{ikt}, the weight of the ML estimate part is increased and the role of the prior is weakened. This desirable specification of MAP estimation makes it appropriate for the cases where the amount of training data is limited. In such cases, the sparsity of the available training data could cause the ML approach to give unrealistic estimates, which can lead to system performance degradation. The MAP estimation equations for other continuous density HMM parameters can also be found in [19] and a somewhat similar interpretation can be given for them.

3. Prior Parameter Estimation

The above discussion obviates the importance of the prior parameters in MAP estimation and their main role in superiority of MAP estimation over ML estimation in sparse data conditions. However, the estimation of appropriate prior parameters in such cases is not trivial. In practical approaches, to overcome this problem, usually, an empirical Bayes approach is followed [20] [21]. Using this approach, several techniques have been adopted to acquire better sets of priors in similar applications (e.g. [22] [23]). However, these techniques usually involve complicated algorithms and need several types of systems to allow parameter calculations. The simpler approach adopted in [24] is found to be sufficient in this case. Here, the parameters are calculated using an already available initial recognition system. As an example, the prior parameters used in (4) are calculated as follows

$$\boldsymbol{\mu}_{ik} = E(\boldsymbol{m}_{ik}) \tag{5}$$

$$\tau_{ik} = \psi_{ik} - 1, \tag{6}$$

where ψ_{ik} represents a parameter of the prior normal-Wishart distribution. Meanwhile, in practice, it was found that a unique value for τ_{ik} would perform equally well. The values for $\boldsymbol{\mu}_{ik}$, however, are the means of the initial recognition system pdfs.

4. Triphone System Construction

Due to the importance of the initial system used for prior parameter calculations, we have tried to construct a better initial system. In the traditional triphone modeling approach, a baseline monophone system is used as the initial system. Then, the parameters of the monophone model of the base phone in any triphone compound are copied as the parameters of that model. This process is often called *cloning*. Hence all the initial triphones with the same middle phone would have the same set of parameters (i.e. all allophones of a single phone). Starting from this point, the parameters of the triphone system are estimated using a maximum likelihood approach. Hence, the cloned triphone system is in fact the same monophone system and further training of the parameters faces the grave problem of training data sparsity. We have proposed an enhanced set of prior parameters, derived from biphone models, as the initial models for MAP estimation of the triphone models.

Consider a context-independent phonetic unit (monophone) is called *u*. Then we call the left context-dependent phonetic unit (biphone) as $l < u$ and the right context-dependent one as $u > r$. In this case, the left and right context-dependent unit (triphone) would be known as $l < u > r$. Ming *et al.* have shown that given a phone-level acoustic observation $\boldsymbol{o}$, applying certain simplifying independence assumptions, one can write [17]

$$p(\boldsymbol{o} \mid l < u > r) = \frac{p(\boldsymbol{o} \mid l < u)\, p(\boldsymbol{o} \mid u > r)}{p(\boldsymbol{o} \mid u)}. \tag{7}$$

This means that triphone models can be replaced by left and right biphones and monophones during likelihood calculations. They have also shown that for an observation sequence of $\boldsymbol{o_1 \ldots o_T}$, the state-based likelihood function can be written as:

$$p(\boldsymbol{o_1} \cdots \boldsymbol{o_T} \mid \lambda) = \sum_{s} \prod_{t=1}^{T} a_{s_{t-1}s_t} \frac{b_{s_t}^{l<u}(\boldsymbol{o_t})\, b_{s_t}^{u>r}(\boldsymbol{o_t})}{b_{s_t}^{u}(\boldsymbol{o_t})}, \tag{8}$$

where λ is the parameter set of the triphone model, s is the occupied state at the time t, a_{ij} is the state transition probability and b_s describes the state occupation probability of the specified state of the specified model.

In (8), the observation probability is found using the observation densities from the states of all three models, i.e. the context-independent one and the two left and right context-dependent models. This approach, although very efficient in memory requirements, needs three probability calculations, in place of one, for every step of the recognition algorithm. As these calculations make up the major part of the computation time during the recognition phase, this might result in high computation costs.

The composed probabilities section of (8), assuming sets of 3-state left-right models, can be written separately for each state as $(b_1^{l<u} b_1^{u>r}) / b_1^u$, $(b_2^{l<u} b_2^{u>r}) / b_2^u$ and $(b_3^{l<u} b_3^{u>r}) / b_3^u$. Then, as a rational simplifying assumption, one can consider the states 2 and 3 in a right biphone model and the states 1 and 2 in a left biphone model as context-independent, i.e. $b_2^{l<u} = b_2^u$, $b_3^{l<u} = b_3^u$, $b_1^{u>r} = b_1^u$ and $b_2^{u>r} = b_2^u$. Hence, the above state probabilities will reduce to $b_1^{l<u}$, b_2^u, $b_3^{u>r}$. In other words, the models of a triphone system can be constructed using appropriate states from the models of monophone, left biphone and right biphone systems.

Taking into account the simplifying assumptions made, the triphone system made using the above proposal will not be an accurate one. Hence, in place of building our main triphone system in the above fashion, we have decided to construct the initial triphone system in this manner so that it can be used later for prior parameter

estimation in a MAP training framework for our triphones. Having only one monophone initial system, we decided to construct two separate left- and right-biphone systems to provide parameters needed for prior parameter calculations.

The procedure followed to construct the triphone system is shown in Fig. 1. All the phone models are assumed 3-state left-right continuous density HMMs. Firstly, 2 biphone model sets were trained independently using all the training data. The biphones were right-context and left-context sets of models respectively and were trained starting from cloned biphones using the parameters of an already available monophone system [25]. The triphone system was synthesized by concatenating the states from the left and right biphones. In this approach, the parameters of the leftmost state were provided by the leftmost state of the left biphone model with the same seed and left context phones. Similarly, the parameters of the rightmost state were provided by the rightmost state of the right biphone model with the same seed and right context phones. Assuming the models to have three states, the parameters of the middle states were replaced by those of the state 2 of corresponding monophone models available in the right biphone set. The reason for this is that as we have used word-internal biphones, at the left word boundaries in the case of left biphones, and at the right word boundaries in the case of right biphones, monophone models appear among the biphones and can be used in the process. Although it seems that the use of monophones from both these model sets would be more appropriate, due to a limitation in our modeling of Persian (Farsi) words, some monophones were not present in the left biphone set. This limitation did not allow a word to start with a vowel, but with a glottal stop preceding the vowel. Note that the glottal stop is also counted as a phoneme in Persian phonetics. Hence, the vowels are absent among the monophone models seen in the left-context biphone model set. Therefore, only the monophone models from the right biphone model set were used.

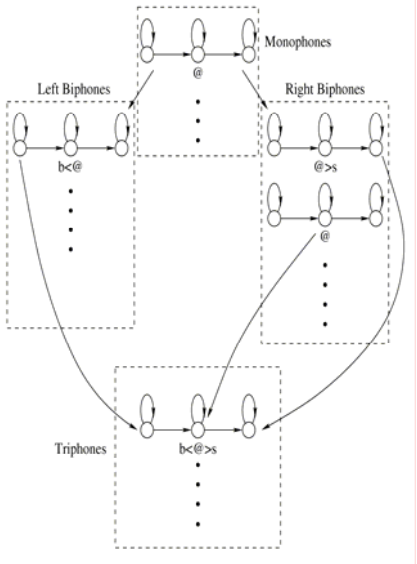

Fig. 1: The procedure followed in synthesizing initial triphones from biphones.

5. Tied Triphone System

The system resulted from the above-mentioned procedure was used to calculate the prior parameters needed for MAP estimation as exemplified by (5) and (6). However, the MAP-estimated system involves a large number of parameters that in turn result in poor parameter estimates. The tying process consists of an initial clustering step which groups the corresponding states of all triphones with similar seed phones (all allophones of the same phone) in separate clusters and a further step to tie the parameters of the states grouped in each cluster. The clustering procedure is a data-driven agglomerative one, similar to those explained in [11] and [10]. This procedure consists of the following steps:

1. Allocate one cluster per state.
2. Find all inter-cluster distances.
3. Find the smallest inter-cluster distance *(d(i,j))*.
4. If *d(i,j)* is not less than *T*, stop.
5. Otherwise, merge clusters *i* and *j* and find all inter-cluster distances with this cluster.
6. Continue from 3.

Here, *T* is a predefined inter-cluster threshold used to control the clustering process. The distances between states were calculated using divergence distance measure. This algorithm was applied to all sets of states chosen as explained and continued until converged. In the end, the state parameters of all states in the same cluster were tied and another phase of training was carried out.

Table 1: Model count and performance evaluation results for monophone and biphone systems.

	Mono-phone System	Left Biphone System	Right Biphone System
Number of Models	32	618	625
Recognition Accuracy (%)	46.9	65.7	64.4

6. Implementation

The above algoriththm was utilized to build a Persian (Farsi) continuous speech recognition system. The system structure was based on a medium-sized vocabulary Persian speech database, FARSDAT [26]. The database was first inspected thoroughly and the utterances from all strong-accented outliers were removed. Then, it was partitioned into training and test sections with a total of about 1800 and 900 sentences respectively, with different speakers in the two sections. The available monophone system [25], consisted of 32 models for 30 basic Persian phonemes plus silence and between-word pause and was built using the above-mentioned training data. All the models, except that of the between-word pause, were 3-state left-right HMMs without skip transitions. The model for the between-word pause was a single-state HMM with a possibility of being bypassed.

The first implementation stage consisted of building untied models for all available word-internal triphones within the database using the monophone models. Firstly, as mentioned in the last section, two biphone systems to include left and right contexts were built. This included cloning all the biphones using their seed monophone models and performing 4 iterations of Baum-Welch re-estimation. This resulted in 618 left and 625 right biphones. Later, the triphones had to be built using the available biphone systems. Note that the total number of available word-internal triphones in the section of the database used in our experiments was 2433.

Before proceeding with the triphone model construction, the performance of both monophone and biphone systems were evaluated. Table 1 summarizes the results obtained. As expected, the biphone models apparently outperformed the context-independent ones in speech recognition evaluations.

The triphone system was built using the technique pointed out in section 3. The resultant system was then used as both the training initial system and the system for calculating prior parameters during the MAP estimation process. Four iterations of MAP estimation were then performed with the MAP parameters calculated at each stage using the results of the last stage. During these MAP estimation processes, the prior parameter $_{ik}$ was set to different values from 2 to 20 for all the states of the system. However, the values of 5 and 10 where found to perform better in this case. Other prior parameters were calculated similar to the approach followed in [24].

The final stage included constructing a tied parameter triphone system using the untied triphone system already available, as pointed out in Section 5. The clustered states were then tied reducing the number of system states from 7297 to 1079, which is less than 15%. Then, 4 iterations of Baum-Welch re-estimation was carried out to further train the tied-state system parameters.

7. Experimental Results

The system was originally implemented as a single-Gaussian system. The reason was that the extension of the system to mixture Gaussians, at any stage, could be carried out by a mixture-incrementing approach and further training of the system parameters.

The speech data in both the training and test sections of FARSDAT corpus were first downsampled from their original sampling rate of 44.1 KHz to 16 KHz and pre-emphasized and blocked into frames of 25 msec. with overlaps of 15 msec. A Hamming window was then applied to the signal and 12 Mel-cepstral coefficients plus log energy were then computed and the delta and delta-delta parameters added to them to extend the total size of the frame feature vectors to 39.

All the results reported in this paper are derived with no language model applied. The reason for this has been the results of our earlier experiments, which showed that the application of a simple language model, such as a word pair grammar, would have led to relatively high recognition rates. This would have prevented us from observing the effects of applying our algorithms to the system. This is found to be due to the artificial nature of the database, which is primarily designed to act as a phonetically balanced database for speech research and not a recognition-specific one [25].

The results of our experiments, together with the result of tree-clustering reported in [27], in similar training and test conditions, are shown in Table 2. Note that the system reported in [27] is the only context-dependent recognition system based on decision tree clustering that has been tested in conditions similar to ours and may be regarded as the Persian counterpart of the system reported in [2]. It is worth remembering that the baseline single-Gaussian monophone system had a no-grammar recognition accuracy of 46.9%. The results are reported as percent recognition accuracy and are given for untied triphones, i.e. the triphone models before the application of a tying procedure, and completely trained tied-state triphones. By ML, we mean models whose parameters are estimated within a Maximum Likelihood framework. The data-driven clustering and tying is one of the usual techniques used for tied-state triphone building where initial models are cloned using monophone models and trained using a ML approach. This approach is equivalent to the approach followed in many references such as [10]. The tying process is exactly the same for IMAP2 approach. However, due to different untied models resulting from ML and MAP approaches, the tied systems are not necessarily identical in shared parameters among different states. The third triphone result belongs to decision-tree based clustering and tying as reported in [2]. This approach also uses the maximum-likelihood approach in initial triphone model building but the clustering approach is different.

Table 2: Comparison of the results obtained by Improved-MAP estimated triphone modeling scheme and the ordinary tied-state triphone modeling.

	Untied Triphones		Tied-state Triphones		
	ML	IMAP1	Data-driven (ML)	IMAP2	Tree-clustered [27]
Recog-nition Accuracy (%)	58.4	61.7	70.2	**72.1**	71.1

As can be seen, the Improved-MAP estimated untied triphone system (named IMAP1) outperforms the similar ML system (ML) by more than 3%. Obviously, the performance of the untied triphone system lags that of both the biphone systems due to much larger number of models, i.e. 2433 versus 618 to 625, leading to under-training for the triphone system.

The MAP estimation result reported is obtained using τ_{ik}=5 for all cases. Furthermore, in spite of the high

capability of MAP estimation in sparse data conditions, the update of parameters is inhibited when a triphone is seen less than 3 times within the training data. This is to prevent unreliable parameter estimates from happening. After tying the parameters and further training the system, the final results still show about 2% improvement for the Improved-MAP-based system (IMAP2) over the ordinary triphone modeling (Data-driven). It also outperforms the widely-used phonetic tree clustering and tying approach.

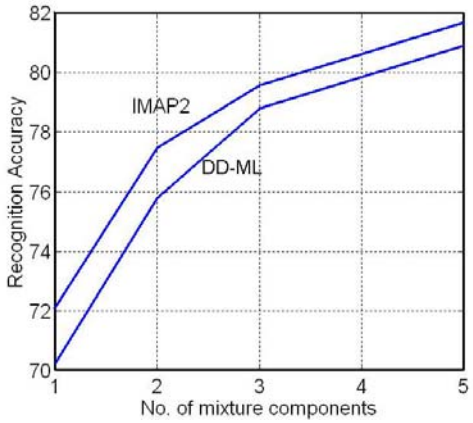

Fig. 2: Comparison of the mixture-Gaussian results obtained from improved-MAP estimated tied-state triphones (IMAP2) and the results of an ordinary ML estimated tied-state triphone system.

Fig. 2 compares the performances of the two data-driven tied-state triphone systems in mixture-Gaussian case and with different numbers of mixture components in output distributions. Once again, the Improved-MAP estimated triphone system (IMAP2) consistently outperforms the ordinary data-driven tied-state system (DD-ML) with increasing numbers of mixture components. It should be noted that further increase in the number of mixture components did not improve the results considerably. This is caused by the limited amount of training data available.

8. Conclusion

The Improved-MAP estimated triphone system was implemented to asses the effect of MAP estimation with improved priors in building a triphone system using data driven tying approach. It was shown that this approach could improve the performance of the resultant triphone system in comparison to ordinary ML-based parameter estimation. The improvement is believed to be due to both better initial modeling and better tying. These are achieved by the use of better-trained untied triphones obtained by Improved-MAP estimation.

Although the approach was applied to a word-internal triphone system, it is believed to work equally well for the case of word-external (cross-word) triphones. Furthermore, since the initial untied parameters are better estimated in this approach, its application to other clustering approaches such as tree-based clustering might also be useful.

Acknowledgment

This research has been funded by national research projects grant no. NRCI 357 and was supported by the Iranian National Research Council.

References

[1] C-H. Lee, J-L. Gauvain, R. Pieraccini and L.R. Rabiner, "Subword-Based Large-Vocabulary Speech Recognition", *AT&T Technical Journal*, pp. 25-36, Sept./Oct. 1993.

[2] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev and S.J. Young, " The Development of the 1994 HTK Large Vocabulary Speech Recognition System," in *Proc. ARPA Spoken Language Technology Workshop*, Bartoncreek, 1995.

[3] S-L. Wu, B.E.D. Kingsbury, N. Morgan and S. Greenberg, "Incorporating Information from Syllable-Length Time Scales into Automatic Speech Recognition," in *Proc. ICASSP-98*, Seattle.

[4] O-W Kwon, "Performance of LVCSR with Morpheme-based and Syllable-based Recognition Units," *in Proc. ICASSP-2000*, Istanbul.

[5] J. Ogata and Y. Ariki, "Syllable-Based Acoustic Modeling for Japanese Spontaneous Speech Recognition," in *Proc. EUROSPEECH-03*, Geneva.

[6] B. Imperl, Z. Kacic, B. Horvat and A. Zgank, "Agglomerative vs. Tree-based Clustering for the Definition of Multilingual Set of Triphones," in *Proc. ICASSP-2000*, Istanbul.

[7] J. Park and H. Ko, "Compact Acoustic Model for Embedded Implementation," *in Proc. ICSLP-04*, Jeju.

[8] K-F. Lee, "Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition," *IEEE Trans. Acoust., Speech, Sig. Proc.*, vol. 38, 1989.

[9] A. Ljolje, "High Accuracy Phone Recognition using Context Clustering and Quasi-triphonic Models," *Computer Speech and Language*, vol. 8, No. 2, 1994.

[10] S.J. Young and P.C. Woodland, "The Use of State Tying in Continuous Speech Recognition,," in *Proc. EUROSPEECH-93*, Berlin, pp. 2203-2206.

[11] S.M. Ahadi, "Reduced Context Sensitivity in Persian Speech Recognition via Syllable Modeling," in *Proc. 8th Australian International Conference on Speech Science and Technology (SST'2000)*, Canberra.

[12] S. Kanokphara, "Syllable Structure Based Phonetic Units for Context-Dependent Continuous Thai Speech Recognition," in *Proc. EUROSPEECH-2003*, Geneva, pp. 797-800.

[13] D. Willett, C. Neukirchen, J. Rottland and G. Rigoll, "Refining Tree-based State Clustering by Means of Formal Concept Analysis, Balanced Decision Trees and Automatically Generated Model-Sets," in *Proc. ICASSP-99*, Phoenix, Arizona, vol.2, pp. 565-568.

[14] K. López de Ipiña, M. Graña, N. Ezeiza, M. Hernández, E. Zulueta and A. Ezeiza, "Decision Tree-Based Context Dependent Sublexical Units for Continuous Speech Recognition of Basque," in *Progress in Pattern Recognition, Speech and Image Analysis*, pp. 259-265, Springer Berlin / Heidelberg, 2003.

[15] J-L. Gauvain, and C-H. Lee, "Bayesian Learning of Gaussian Mixture Densities for Hidden Markov Models," *Speech Communication*, vol. 11, 1992, pp.205-213.

[16] P. Somervuo, "Comparison of ML, MAP, and VB based Acoustic Models in Large Vocabulary Speech Recognition," in *Proc. ICSLP-04*, Jeju.

[17] J. Ming, P. O'Boyle, M. Owens and F.J. Smith, "A Bayesian Approach for Building Triphone Models for

Continuous Speech Recognition," *IEEE Trans. Speech and Audio Proc*., vol. 7, No. 6, Nov. 1999, pp. 678-684.
[18] F. Brugnara, "Model Agglomeration for Context-Dependent Acoustic Modeling," in *Proc. EUROSPEECH-2001*, Aalborg, Denmark, pp. 1641-1644.
[19] J-L. Gauvain and C-H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. Speech and Audio Proc*., vol. SAP-2, No. 2, April 1994, pp.291-298.
[20] H. Robbins, "The Empirical Bayes Approach to Statistical Decision Problems," *Ann. Math. Stat.*, Vol. 35, pp.1-20, 1964.
[21] J.S. Maritz and T. Lwin, *Empirical Bayes Methods*, 2nd Ed., Chapman and Hall, 1989.
[22] Q. Huo and C. Chan, "Bayesian Adaptive Learning of the Parameters of Hidden Markov Models for Speech Recognition," Technical Report, Department of Computer Science, University of Hong Kong, September 1992.
[23] S.M. Ahadi, "On Prior Parameter Estimation for Bayesian Adaptation of Continuous Density Hidden Markov Models," *Amirkabir Journal of Science and Technology*, vol.10, No.39, Autumn/Winter 1998/99.
[24] C-H. Lee and J-L. Gauvain, "A Study on Speaker Adaptation for Continuous Speech Recognition," in *Proc. ARPA Cont. Speech Rec. Workshop*, September 1992, Stanford, pp. 59-64.
[25] S.M. Ahadi, "Recognition of Continuous Persian Speech Using a Medium-sized Vocabulary Speech Corpus," in *Proc. EUROSPEECH-99*, Budapest.
[26] M. Bijankhan *et al.*, "FARSDAT - The Speech Database of Farsi Spoken Language," in *Proc. 5th Australian International Conference on Speech Science and Technology (SST'94)*, Perth.
[27] S.H. Shams and S.M. Ahadi, "Context Dependent Modeling in Continuous Speech Recognition Based on a Persian Phonetic Decision Tree," *Modarres Technical and Engineering Journal*, vol. 13, Fall 2003.