ACOUSTIC MODELING BY PHONEME TEMPLATES AND MODIFIED ONE-PASS DP DECODING FOR CONTINUOUS SPEECH RECOGNITION

V. Ramasubramanian[1], *Kaustubh Kulkarni*[1], *Bernhard Kaemmerer*[2]

[1] Professional Speech Processing - India, Siemens Corporate Technology, Bangalore, India
[2] Professional Speech Processing - Munich, Siemens Corporate Technology, Munich, Germany
{V.Ramasubramanian,Kulkarni.Kaustubh,Bernhard.Kaemmerer}@siemens.com

ABSTRACT

We propose a novel framework for continuous speech recognition (CSR) based on non-parametric acoustic modeling using multiple phoneme templates set in a modified one-pass DP decoding algorithm, in contrast to the conventional HMM acoustic models set in Viterbi decoding. We particularly emphasis the 'selectivity' property of templates as set in the proposed modified one-pass DP decoding algorithm and explore various contextual definitions of the templates and their relative performances for a range of small vocabulary tasks with TIMIT database using only acoustic models. Based on this, we show that the proposed framework based on phoneme template modeling is a viable means for CSR with potential for interesting issues in acoustic modeling and decoding strategies, particularly in the paradigmatically novel framework of model-free CSR.

Index terms: Continuous speech recognition, Phoneme template modeling, One-pass DP decoding

1. INTRODUCTION

Current continuous speech recognition (CSR) is largely based on HMM based acoustic modeling of phones and triphones. However, various shortcomings of HMMs have long been felt now, mainly with respect to its inability to account for inter-frame correlations and the difficulties in reliably estimating very large number (millions) of context-dependent HMM parameters from limited training data. In keeping with these long felt needs to overcome the shortcomings of HMM, there have been some attempts in the literature to explore more efficient alternatives, with varied degrees of success, including the stochastic segment models, trajectory models etc., also in parametric forms.

In a very recent work, De Wachter et al. [2] propose the use of template based continuous speech recognition using a template database. Here, the main approach is to use a template database of continuous speech which is annotated phonetically with various acoustic as well as non-verbal attributes. During decoding, the system uses a token passing strategy to search this continuous unit database for an optimal word-level decoding by employing a DTW based matching between the input utterance (feature vector sequence) and the templates present in the continuous speech database (template database). This search recoveres the optimal template sequence (and decoded word sequence) in the continuous template database which best matches the input speech, by assigning template concatenation costs in correspondence to template transitions allowed by the lexicon of the words of the task, thereby ensuring that the 'unit-selection' type of decoding benefits from presence of consecutive templates in the continuous template database, but conforming to the lexicon of the words in the vocabulary.

In contrast, in this paper, we propose acoustic modeling by use of multiple templates of a monophone (context-independent phones) or triphones (context-dependent phones) drawn from the unit database constituting the training data; these multiple templates are drawn out of the training data and kept as the phoneme inventory in keeping with the de facto CSR framework, with the only difference of replacing an HMM phone (monophone or triphone) model by a set of multiple templates (referred here as 'phoneme template models'). These are incorporated in a CSR framework, where the decoding is done by a modified one-pass dynamic programming (DP) algorithm requiring more complex recursions when compared to the conventional one-pass DP algorithm used for connected word recognition [3] (with whole word templates). The pronunciation dictionary (word lexicon) is specified as a linear baseform of phones / triphones as in conventional CSR.

Our algorithm conforms to the basic definition of CSR and does not give any particular emphasis to the natural ordering of the templates in the training data (from which the acoustic models - non-parametric phoneme template model in this case - are obtained) since this would not be of relevance to an arbitrary vocabulary of words of a given task, which in principle can be very different from the ones in the training data. Instead, our emphasis lies on acoustic modeling alone, particularly with respect to the 'selectivity' property of templates as set in the proposed modified one-pass DP decoding algorithm for continuous speech recognition and its implications on acoustic modeling by monophones or triphones, particularly defined in a long-context or short-context, leading to strong possibilities of very high performance with acoustic modeling alone using phoneme templates with long-span contexts.

We do not particularly focus on large vocabulary performances, but instead present results of the proposed CSR framework for a range of small vocabulary sizes with the TIMIT database, since our emphasis is on establishing the importance of the context-span for acoustic modeling with phoneme templates. Based on this, we show that the proposed framework based on phoneme template modeling is a highly viable means for CSR and opens up several interesting issues in acoustic modeling and decoding strategies, particularly in the paradigmatically novel framework of model-free continuous speech recognition.

2. PROPOSED PHONEME-TEMPLATE BASED CONTINUOUS SPEECH RECOGNITION

Fig. 1 shows the main algorithmic framework for continuous speech recognition (CSR) proposed here. The proposed phoneme template based acoustic model and the modified one-pass DP based decoding algorithm are outlined below.

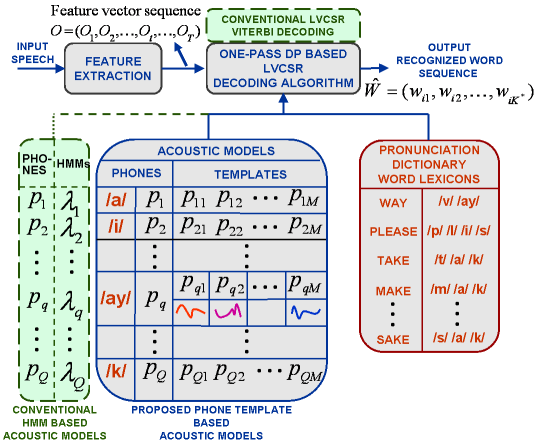

Fig. 1. *Proposed phoneme template based CSR framework*

2.1. Proposed phoneme template modeling

In the proposed phoneme template based acoustic modeling, each phone / triphone is represented non-parametrically by a set of multiple templates drawn from a large training database which is phonetically transcribed. These are shown in Fig. 1 in the block marked 'acoustic models' (solid lines) as $\mathcal{P} = (p_1, p_2, \ldots, p_q, \ldots, p_Q)$, where each phone p_q has up to M templates given by $(p_{q1}, p_{q2}, \ldots, p_{qm}, \ldots, p_{qM})$. The conventional phone-HMM based acoustic models are shown alongside in the block with dashed lines. The trajectory representation of such phone templates are shown in Fig. 2 for the phones /p/ and /l/ (as part of the word 'please' - /p/ /l/ /i/ /s/) to highlight the contextual differences in these templates. The other two phones /i/ and /s/ are given in various triphone forms in Table 1.

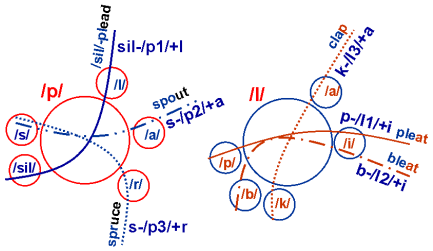

Fig. 2. *Multiple phoneme template trajectories of /p/ and /l/*

Table 1. *Phoneme templates of /i/ and /s/ with triphone contexts*

Phone	From word	Phone	From word
$n-\mathbf{i}_1+d$	*n***ee***d*	$i-\mathbf{s}_1+m$	*pri***s***m*
$l-\mathbf{i}_2+s$	*l***ea***se*	$a-\mathbf{s}_2+p$	*cla***s***p*
$r-\mathbf{i}_3+z$	*fr***ee***ze*	$i-\mathbf{s}_3+sil$	*grea***s***e-sil*

2.2. Proposed decoding algorithm

The use of phoneme template modeling naturally leads to a modified form of the one-pass DP algorithm for CSR. With the continued use of the word models in terms of linear baseform lexicon of the phones / triphones, the one-pass DP decoding algorithm now has to deal with word models composed of sequence of phones / triphones (as in a word's lexicon), each of which is a set of large number of multiple phoneme templates. Fig. 1 shows the proposed modified one-pass DP decoding in the block (solid lines) which now takes the place of the conventional Viterbi decoding block of phone HMM based system (dotted lines).

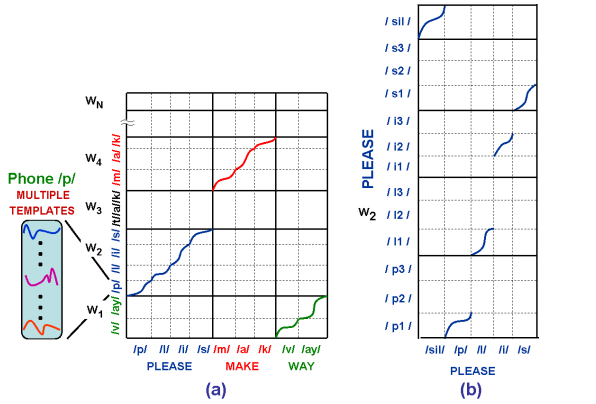

Fig. 3. *(a) Continuous speech decoding in the proposed phoneme template based framework and b) monophone template selectivity and implicit context-dependent modeling*

2.3. Context modeling with phoneme template modeling

Fig. 3(a) and (b) show how the proposed CSR framework uses the phoneme templates for a correct word decoding of continuous speech. Here, each word of the vocabulary is expanded into its phone / triphone sequence using the word's lexicon, and multiple templates of these phones / triphones are used in the y-axis; the input test utterance to be decoded (the continuous speech utterance 'please make way') is shown in the x-axis. Fig. 3(b) shows the expansion of the word-lexicon of 'please', where the multiple templates of the monophones of the word 'please' have three different context-dependencies as shown in Fig. 2 and Table. 1. Further, Fig. 3(b) shows how the various acoustic segments (phones) of the 'please' part of the input utterance are 'selectively' matched to the corresponding 'best' monophone template of the respective phones (from y-axis) by the optimal path (composed of the individual phone warping paths) derived by the proposed one-pass DP algorithm.

This 'selectivity' property of the one-pass DP algorithm has important implications in modeling various context-dependencies as follows. The multiple templates of a phone can be drawn from a very large training data which contains all possible signatures of that phone (such as context, and 'other' acoustic attributes such as speaker, accent, dialect, prosody, environmental noise etc.). Now, an acoustic segment of the input speech utterance, as typified by a particular left/right context (or a speaker, or accent, or dialect or prosody or a specific environmental noise), can be matched to a particular template of the phone that 'best matches' the input acoustic segment. Such a 'selectivity' property provides a mechanism for high classification accuracies for input test data with such intrinsic high variabilities; this is a property unique to non-parametric acoustic modeling, which is not possible with phone-HMM models which lack such 'specificity' due to its parametric form.

In addition to such 'selectivity' properties of templates in general, there are specific advantages of using 'triphone templates' of the triphone contexts of phones in a word, rather than use all possible templates of a monophone occurring in the word. By this, the triphone templates continue to retain the 'selectivity' property for the 'other' acoustic attributes with the advantage of representing the triphones of a word with far less number of triphone templates than monophone templates.

In this paper, we study the use of both monophone and triphone templates; the phoneme template acoustic models in Fig. 1 are treated as monophones or triphones accordingly.

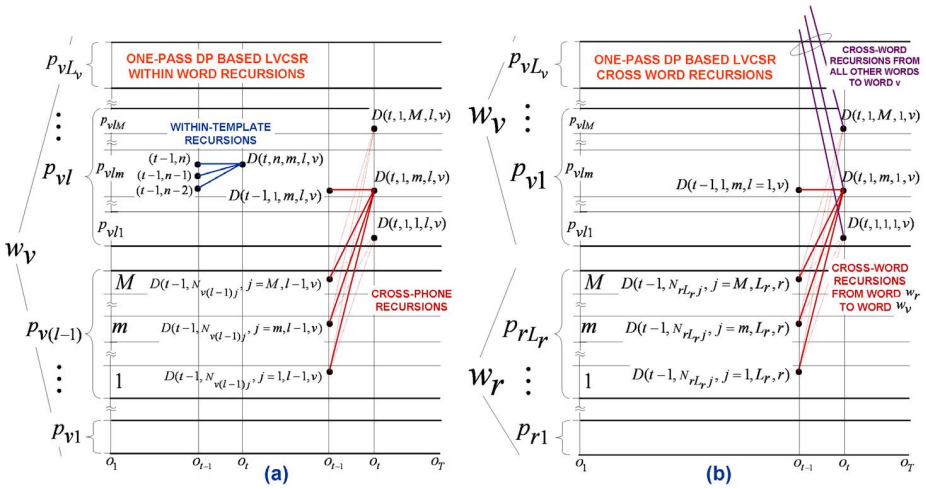

Fig. 4. *Proposed one-pass DP based CSR using multiple phoneme templates: (a) within word and (b) cross-word recursions*

3. PROPOSED ONE-PASS DP BASED CSR

In this section, we describe the proposed one-pass DP algorithm for continuous speech decoding in detail. This has major modifications in its recursions (with respect to conventional connected word recognition) [3], to account for word representations by linear baseform lexicon as in CSR, i.e. as a sequence of phones / triphones each with multiple templates. An example decoding of this algorithm was shown in Fig. 3.

Let the vocabulary words be $\mathcal{W} = (w_1, w_2, \ldots, w_v, \ldots, w_V)$. Each word w_v has a linear baseform lexicon of L_v phones (or triphones) given by $(p_{v1}, p_{v2}, \ldots, p_{v(l-1)}, p_{vl}, \ldots, p_{vL_v})$. Each phone (or triphone) p_{vl} is some phone from the phone/triphone acoustic models $\mathcal{P} = (p_1, p_2, \ldots, p_q, \ldots, p_Q)$, where each phone p_q has up to M templates given by $(p_{q1}, p_{q2}, \ldots, p_{qm}, \ldots, p_{qM})$. Thus, the l^{th} phone in the lexicon of word w_v will have up to M templates given by $(p_{vl1}, p_{vl2}, \ldots, p_{vlm}, \ldots, p_{vlM})$. Each of these templates p_{vlm} has N_{vlm} frames.

These notations and the one-pass DP recursions for CSR using multiple phone/triphone templates are shown in Fig. 4. In all the different recursions, the one-pass DP calculates the optimal (minimum) accumulated distortion $D(t, n, m, l, v)$ to reach the n^{th} frame of template m of the phone / triphone in the l^{th} position of the lexicon of word v, at every time instant $t = 1, \ldots, T$ of the input continuous speech utterance. The local distance $d(t, n, m, l, v)$ in these recursions is the distance between the t^{th} frame of the input speech and the n^{th} frame of template m of the l^{th} phone in the lexicon of word v.

The recursions of the proposed one-pass DP algorithm for continuous speech decoding are as follows.

1. Within-word recursions

a) Within-phoneme-template recursion: These recursions are applied for each of the multiple templates of each phone in the word-lexicon; within a phone-template, these are applied for all frames that are not phone-template-beginning frames, i.e., calculate $D(t, n, m, l, v)$ only for the template-interior frames $n = 2, \ldots, N_{vlm}$, for all templates $m = 1, \ldots, M$ of all the phones in the lexicon $l = 1, \ldots, L_v$ of a word v, for all words $v = 1, \ldots, V$ and for every time instant $t = 1, \ldots, T$.

$$D(t, n, m, l, v) = d(t, n, m, l, v) + \min_{j=(n, n-1, n-2)\&(j>0)} [D(t-1, j, m, l, v)] \quad (1)$$

b) Cross-phone recursion: This is defined for transition from any of the M multiple templates of phone $p_{v(l-1)}$ to any of the M multiple templates of phone p_{vl} in the lexicon of a word w_v. This is applied for all phones excluding the first phone in the lexicon of a word (as it can receive a transition only from other words). These recursions correspond to entry into any of the M templates of a phone within a word, i.e., calculate $D(t, n = 1, m, l, v)$ for $n = 1$ and $m = 1, \ldots, M$ for every phone $p_{vl}, l = 2, \ldots, L_v$ for all words $v = 1, \ldots, V$ and at every time instant $t = 1, \ldots, T$.

$$D(t, n=1, m, l, v) = d(t, n=1, m, l, v) + \min[D(t-1, n=1, m, l, v), \min_{j=1,\ldots,M} [D(t-1, N_{v(l-1)j}, j, l-1, v)]] \quad (2)$$

2. Cross-word transitions

These transitions correspond to entry into the first frame $n = 1$ of any of the M multiple templates $m = 1, \ldots, M$ of the first phone $p_{v(l=1)}$ of the word v from any of the M multiple templates $m = 1, \ldots, M$ of the last phone p_{rL_r} of all words $r = 1, \ldots, R$, i.e., calculate $D(t, n = 1, m, l = 1, v)$ for every time instant $t = 1, \ldots, T$, for $n = 1, m = 1, \ldots, M, l = 1$ and $v = 1, \ldots, V$ as,

$$D(t, n=1, m, l=1, v) = d(t, n=1, m, l=1, v) + \min[D(t-1, n=1, m, l=1, v), \min_{r=1,\ldots,V} [\min_{j=1,\ldots,M} D(t-1, N_{rL_r j}, j, L_r, r)]] \quad (3)$$

3. Termination and backtracking

The decoding using the above recursions terminates at the last time instant T with the optimal accumulated distortion D^*,

$$D^* = \min_{v=1,\ldots,V} \min_{m=1,\ldots,M} D(T, N_{vL_v m}, m, L_v, v) \quad (4)$$

i.e., this is the minimum accumulated distortion over the last frames $N_{vL_v m}$ of all the M templates of the last phone L_v of all the words $v = 1, \ldots, V$. The optimal path through the one-pass DP 'grid' and the corresponding decoded word sequence are recovered by backtracking using the backpointers stored during the forward computations of the recursions given by Eqns. (1), (2) and, (3). The backpointer equations and the backtracking equations are not shown here due to space constraints.

4. EXPERIMENTS AND RESULTS

We evaluate the proposed CSR framework with phoneme templates using the TIMIT database. The experiments done here are primarily intended to bring out the acoustic modeling efficacy of phoneme templates in various contextual settings, rather than on the largeness of the continuous speech recognition tasks or the use of language models, efficient search etc. Keeping with the fact that the proposed system is a template based system, we have used conventional connected word recognition using whole-word templates [3] as the baseline for our comparisons to represent the ideal long-span context for template based modeling.

We have performed 5 main continuous speech recognition experiments using the TIMIT database. Of these, 4 are defined on a task of 21 word vocabulary from the two sentences (sa1, sa2) of TIMIT; the test sentences comprise 100 sentences made of these (sa1, sa2) sentences spoken by 50 speakers. Table 2 shows the definition of the phoneme templates for each of these 4 experiments, where a phoneme template type could be a mono-phone (context-independent) or triphone templates and further defined as out-of-word-context and in-word-context.

Table 2. *Experiments (1-4) & their phoneme template definitions*

Template Type	**Templates drawn from**	
	Outside (sa1,sa2)	Inside (sa1, sa2)
Monophone	1. Mono-non-sa1-sa2	3. Mono-in-sa1-sa2
Triphone	2. Tri-non-sa1-sa2	4. Tri-in-sa1-sa2

Expt. 5 is to show the performance with respect to vocabulary size, with test data comprising of (non-sa1-sa2) sentences in TIMIT with vocabulary words ranging from 20 to 100 words. This is a typical CSR scenario, where triphone templates are drawn from training data which does not have any of the words in the test data.

In all these experiments, the data from which the phoneme templates are drawn are spoken by speakers different from the test speakers and the lexicon used for the words in the vocabulary are as given by the manual phonetic transcription associated with these words in the test sentences. The features used are MFCCs of dimension 39 with delta and delta-delta coefficients and normalized energy.

We present results in terms of word recognition accuracies as computed in CSR decoding between the reference and decoded word sequences; results of Expt. 1 to 4 are shown in Fig. 5(a) and Fig. 5(b) shows the results of Expt. 5.

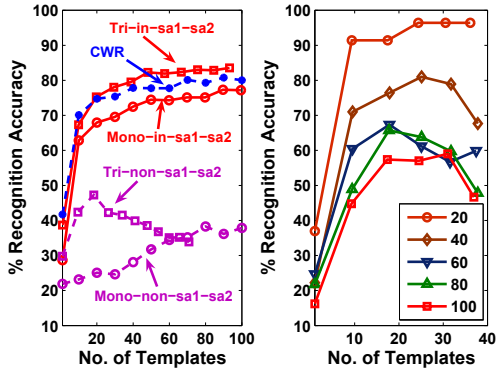

Fig. 5. *Word recognition accuracies (%) of proposed phoneme template modeling; (a) Expts. 1 to 4 (b) Expt. 5*

The following can be noted from Fig. 5(a) with respect to Expts. 1 and 2: (i) The recognition accuracy for the context-independent monophone templates for non-sa1-sa2 cases increases significantly with the number of monophones, but soon saturates with about 80 monophones. This is largely due to its ability to model all possible associated context-dependencies as well as inter-speaker variabilities present in the test utterances by the 'selectivity' property. (ii) The use of triphone-non-sa1-sa2 has 10% (abs) higher performance than monophones from non-sa1-sa2, clearly indicating the advantage of using context-dependent templates. (iii) A much smaller number ($\tilde{2}0$) of triphones are needed to achieve the same (and even better) performances of a larger number (80-100) of monophones.

The following can be noted from Fig. 5(a) with respect to Expts. 3 and 4: The recognition accuracies using monophones and triphones from within the sa1 and sa2 words (mono-in-sa1-sa2 and tri-in-sa1-sa2) are 77-83%, dramatically higher by 30-40% (absolute) than for phoneme templates from non-word contexts in Expt. 1 and 2. This clearly shows the importance of preserving longer contexts in the templates. The triphone performance in-sa1-sa2 itself is about 10% higher than for mono-in-sa1-sa2. This has important implications that use of carefully selected, 'long' context phoneme templates can lead to high accuracies using acoustic modeling alone.

Fig. 5(a) also shows the baseline whole-word based connected word recognition (CWR) [3] as a comparison with our proposed system. It can be noted that the triphone performances are better than CWR with the same number of whole word templates. This has the important implications that the performance of a whole word template can be reached with the use of long-span template models and that a given number of phoneme templates handles larger variability in the test data than the same number of whole word templates.

From the results of Expt 5 in Fig. 5(b), we observe the natural trend of decreasing recognition accuracies from 96% to 57% with increase in vocabulary size from 20 to 100, for a given number of triphone templates. In all the experiments we have reported here, we have focused on the efficacy of only the acoustic model using phoneme template modeling. The recognition accuracies will tend to increase for larger vocabularies with the use of more triphone templates as well as carefully optimized triphone (and longer-span template) categories with decision tree methodologies. It should be noted that the recognition accuracies of 83% in Fig. 5(a) and 57-96% in Fig. 5(b) are obtained with acoustic-modeling alone and are comparable to HMM based acoustic-model-only performances, as was reported earlier in [1] and in more recent state-of-the-art systems.

5. CONCLUSION

We have proposed a novel framework for continuous speech recognition (CSR) based on non-parametric acoustic modeling in terms of multiple phoneme templates (monophones/triphones) set in a one-pass DP decoding framework modified for continuous speech recognition. Based on the results we have obtained for small vocabulary continuous speech recognition using phoneme template acoustic modeling, the proposed framework appears as a viable means for CSR with potential new ways for handling the acoustic units with a rich interplay between the acoustic models and the decoding algorithm and towards alternate CSR frameworks with the advantages of non-parametric acoustic modeling.

6. REFERENCES

[1] K. F. Lee. *Automatic speech recognition – The development of the SPHINX system*. Kluwer Academic, Boston, 1989.

[2] M. De Wachter, M. Matto, K. Demuynck, P. Wambacq, R. Cools and D. Van Compernolle. Template-based continuous speech recognition. In *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1377-1390, vol. 15, no. 4, May 2007.

[3] H. Ney. The use of one-stage dynamic programming algorithm for connected word recognition. *IEEE Trans. on Acoust., Speech and Signal Proc*, 32(2):263–271, Apr 1984