LATENT PERCEPTUAL MAPPING WITH DATA-DRIVEN VARIABLE-LENGTH ACOUSTIC UNITS FOR TEMPLATE-BASED SPEECH RECOGNITION

Shiva Sundaram [1], *Jerome R. Bellegarda*[2]
[1]Deutsche Telekom Laboratories, Ernst-Reuter-Platz-7, Berlin, GERMANY
[2]Apple Inc., 3- Infinite Loop, Cupertino, California, USA.
shiva.sundaram@telekom.de, jerome@apple.com

ABSTRACT

In recent work, we introduced *Latent Perceptual Mapping* (LPM) [1], a new framework for acoustic modeling suitable for template-like speech recognition. The basic idea is to leverage a reduced dimensionality description of the observations to derive acoustic prototypes that are closely aligned with perceived acoustic events. Our initial work adopted a bag-of-frames strategy to represent relevant acoustic information within speech segments. In this paper, we extend this approach by better integrating temporal information into the LPM feature extraction. Specifically, we use variable-length units to represent acoustic events at the supra-frame level, in order to benefit from finer temporal alignments when deriving the acoustic prototypes. The outcome can be viewed as a generalization of both conventional template-based approaches and recently proposed sparse representation solutions. This extension is experimentally validated on a context-independent phoneme classification task using the TIMIT corpus.

Index Terms: latent perceptual mapping, template-based speech recognition, acoustic modeling, data-driven speech units, dimensionality reduction

1. INTRODUCTION

Approaches to acoustic modeling for automatic speech recognition can be broadly categorized into template-based and statistical methods. Template-based systems operate by directly comparing the utterance to be recognized with training instances known as *templates*. They were common in the early days of speech recognition, but as recognition tasks grew in perplexity, their computational requirements became prohibitive. Over the years, they have largely been replaced by statistical systems that predominantly use hidden Markov models (HMM).

While the HMM framework leads to efficient algorithms for training and recognition, it also entails a loss of acoustic information via, e.g., the smoothing effects caused by inexact parametric distributions, which are known to be deleterious for recognition [2]. Increasing the number of modeling parameters mitigates this loss, but like template-based methods that requires significantly more computations to estimate the parameters [3]. Approaches like [4] by Axelrod *et al.* and [5] by Zhao *et al.* use dynamic time warping (DTW) to compensate for the model smoothing, while other approaches forgo the HMM framework entirely in favor of other statistical techniques [6, 7, 8]. With the steady increase in available computational power in recent years, interest in template-like methods [3, 9] has regained momentum as well.

In previous work, we introduced an alternative acoustic modeling framework called latent percepual mapping (LPM) [1] which adopts a similar premise. It offers a template-like solution to acoustic modeling for speech recognition, with the particularity that the acoustic information is derived from weighted frequency counts between suitable data-driven acoustic units and associated speech segments. The framework is inspired by latent semantic analysis in information retrieval [10], except that here *documents* are segments of speech and *units* are entries in a codebook that suitably encapsulates the acoustic feature space. Phoneme classification is thus performed in a *latent* feature space after reducing the representational dimension using singular-value decomposition (SVD). This information extraction procedure shares its motivation with histogram of acoustic co-occurrence (HAC) models by van Hamme [11], where the latent structure in speech utterances is decomposed by finding repeated acoustic patterns.

In this paper, we generalize the LPM framework by contrasting fixed and variable-length units for building the unit-document matrix. Experimental results on the TIMIT acoustic-phonetic corpus illustrate the relationship between (generalized) LPM and conventional template (frame- or DTW-based) solutions. They also serve as backdrop to discuss links between LPM and recent efforts exploiting sparse representations. The paper is organized as follows. In the next section, we detail how to derive a suitable set of variable-length acoustic units and integrate it into the LPM paradigm using DTW. Section 3 discusses the parallels with sparse representations and other template-based approaches. In Section 4, we describe the experimental setup considered. Finally, Section 5 illustrates the performance of the proposed approach on the TIMIT context-independent phoneme classification task.

2. LPM WITH VARIABLE-LENGTH UNITS

LPM establishes an analogy between text documents (made up of words) and speech segments (composed of suitable acoustic units) [1]. In the same way that words span a well-defined vocabulary, acoustic units span a discrete set obtained by vector quantization of the underlying speech feature vectors. The LPM approach is a template-based framework for speech recognition; no prior model for the phoneme segments is assumed and the steps followed to obtain prototypical templates are purely data-driven. The training and recognition steps illustrated in Figure 1 are described below.

Training comprises of three main steps: (1) extracting relevant *units* from a given set of phoneme instances; (2) deriving a unit-document co-occurrence matrix; and (3) mapping the phoneme instances to a dimensionality reduced latent space after singular value decomposition (SVD) of the co-occurrence matrix.

1. **Derivation and selection of units:** We assume M phoneme segments are available for training. A codebook $\mathcal{C}_N$ (with N entries) is first derived after feature extraction by frame-based analysis. The codebook is then used to vector quantize (VQ) the feature-vectors in the training set. All, 1-gram, 2-gram and 3-gram quantized sub-sequences in the phoneme instances are considered together. Furthermore, based on our results in [1] and other phoneme classification work, we also separately considered sub-sequences obtained by partitioning phoneme segments in 30-40-30% sub-sequences. Note that irrespective of the length of the sub-sequences, we take a *bag-of-acoustic-units* approach instead of a *bag-of-features*

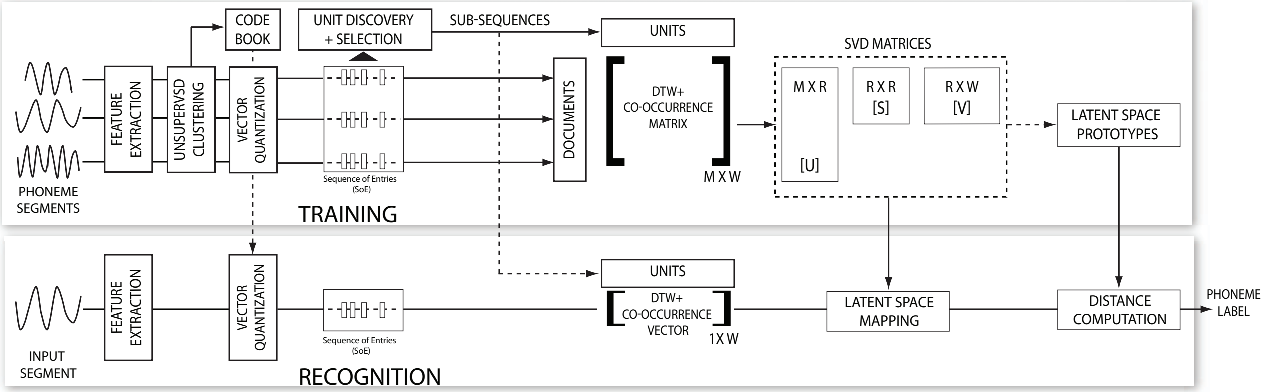

Fig. 1. Latent Perceptual Mapping using Sub-sequences for Phoneme Classification.

approach outlined in [1]. The LPM approach introduced in [1] is in fact a special case of this approach where only units of length 1 (1-gram) are considered.

Even for a reasonable codebook size N, the number of units can be quite large. Some units (sub-sequences) can repeatedly appear in a large number of phoneme instances while others may only appear in a handful of them. In this work, we investigated different ranking methods to select a subset of those units and show that it is indeed possible to have comparative performance with reduced dimensional representation in the latent space with lesser number of units. The ranking methods use two empirical measures: (1) the indexing power (Φ^i) and (2) the empirical probability (π) of unit given phoneme class. After ranking, top W units $\mathcal{W}$ are selected for calculating the unit-document co-occurrence matrix and the subsequent steps.

Indexing Power: The indexing power Φ^i of a unit is given by $\Phi^i = (1 - \varepsilon_i)$ where ε_i is the empirical entropy measure of the i^{th} unit,

$$\varepsilon_i = \frac{-1}{logM} \sum_{m=1}^{m=M} \frac{\kappa_{i,m}}{\tau_i} log(\frac{\kappa_{i,m}}{\tau_i}) \tag{1}$$

Here $\kappa_{i,m}$ is the number of times the i^{th} unit appears in the m^{th} phoneme instance and $\tau_i = \sum_{\forall m} \kappa_{i,m}$ is the total number of times the i^{th} unit appears in the collection complete collection of training segments.

Empirical Probability: The empirical probability π_p^i of i^{th} unit in a phoneme class p is given by

$$\pi_p^i = \frac{\kappa_i^p}{\lambda_p} \tag{2}$$

where κ_i^p is the number of times the i^{th} unit appears in instances of phoneme p and λ_p is the total number of instances in phoneme p.

The indexing power and the empirical probability favor longer less frequently occurring units and shorter more frequently occurring units respectively. Using the two empirical measures it therefore possible to derive different ranks units in the training instances and control the selection of units accordingly. The details of these are described later in section 4.

2. **Co-occurrence Matrix:** A $M \times W$ unit-document co-occurrence matrix F is calculated by counting the number times each unit in $\mathcal{W}$ appears in the m^{th} phoneme instance $\forall m$. The $(m, w)^{th}$ entry of F is obtained as follows:

$$f(m, w) = \left(\frac{\sum_{j \in \mathcal{A}^m} I_w(j)}{\lambda_m} \right) \cdot p_w, \tag{3}$$

$$\text{where } w \in \{1, 2, \ldots, W\}.$$

Here, λ_m is the total number of units in $\mathcal{A}^m$ (the phoneme instance). The indicator function $I_w(j) = 1$ iff the w^{th} unit in $\mathcal{W}$ is nearest to the j^{th} unit in $\mathcal{A}^m$ belonging to the m^{th} phoneme instance. As each unit is a sequence of vectors, dynamic time warping (DTW) with appropriate length normalization [4] is used to determine $I(\cdot)$.

3. **Dimensionality reduction:** After obtaining F, a reduced-rank approximation of the matrix F can be obtained by SVD as follows:

$$\hat{F}_{M \times W} = U_{M \times R} \cdot \Sigma_{R \times R} \cdot V_{W \times R}^T \tag{4}$$

Here U and V are the left and right singular vector matrices and Σ is the diagonal matrix of singular values. R essentially approximates the rank of F ($R \leq \min(M, W)$). The m^{th} segment in the collection is characterized by f_m, the i^{th} row of F. From (4), f_m can be projected onto the orthornormal basis formed by the row vectors of V^T, or, equivalently, the column vectors of V. This defines a new representation, namely $u_m \cdot \Sigma$, where u_m is the m^{th} row of U. In essence, the row vector:

$$y_m = u_m \cdot \Sigma \tag{5}$$

characterizes the position of the segment m in the underlying (R-dimensional) latent space. In the LPM framework, phoneme segments from the training set are mapped to vectors in the latent space and then used as acoustic prototypes. As shown in Figure 1, obtaining the units and the resulting latent perceptual space constitutes the LPM training procedure.

Classification of an unknown test segment is performed by following the "folding in" approach illustrated in [12]. Mapping a new segment not belonging to the original collection M is straightforward using $\mathcal{W}$ and the singular vectors obtained during training. A new (test) segment x can be represented as:

$$y_x = u_x \cdot \Sigma = f_x \cdot V \tag{6}$$

where f_x is the unit-document feature vector associated with the new segment (which is treated as an additional row of the matrix F).

The similarity between two segments is obtained by a dot product between the associated vectors in the latent space. A nearest-neighbor rule is then adopted to predict the phoneme label of the unknown test segment.

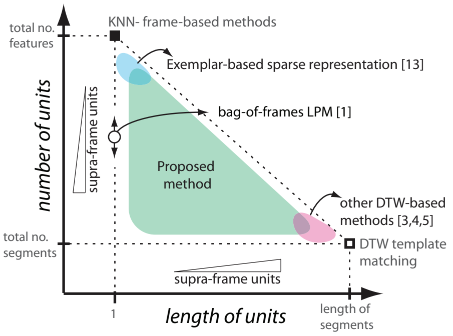

Fig. 2. LPM in relation to other template-based approaches.

3. LINK WITH OTHER TEMPLATE-BASED METHODS

The extension to the original LPM approach described in the previous section involves integrating temporal information in an unsupervised manner directly into the LPM feature extraction. Specifically, we represent acoustic events via unsupervised sequences of supra-frame units and rely on supra-frame distance metrics such as dynamic time warping (DTW) for deriving the unit-document matrix in LPM. In this way, the LPM framework emerges as a generalization of conventional DTW-based methods. In fact, by further increasing the length of the units (to, say, complete phoneme segments), the LPM framework leads to a DTW-like template comparison approach.

This aspect is illustrated in the bottom right corner of Figure 2. This diagram depicts how LPM relates to various template-based methods via two parameters, the number of units and the length of the units. On Figure 2 the x-axis displays the possible length of units expressed in frames, which varies from 1 (in which case each frame/feature vector is its own unit) to the length of complete phoneme segments (assuming phoneme boundaries are known). The y-axis displays the possible number of units, which varies from the number of segment instances available to the total number of features extracted. For phoneme classification, DTW-based methods operate based on the full length of all observed exemplars.

In contrast, in our original study [1], where units were frames, we varied the number of units by unsupervised K-means clustering of the feature vectors. By using as many units as the number of features, the LPM framework essentially morphs into a frame-based K-nearest neighbor (KNN) method. As the number of units decreases, the representation becomes more parsimonious. This aspect is illustrated on the left hand side of Figure 2, which shows bag-of-frames LPM moving vertically along the $x = 1$ line.

Finally, the top-left corner of the figure illustrates the relationship with exemplar-based classification using sparse representations (EBCSR) recently proposed by Kanevsky et al. [13]. As with KNN, *all* features are presented to the classification mechanism for training, but in EBCSR, each test observation vector y is represented as a weighted sum of T d dimensional training examples collected in a matrix $H_{d\times T}$, i.e., $y = H \cdot \beta$, where the weight vector β satisfies some suitable sparseness constraint, and therefore has only a few non-zero elements. In other words, it is assumed that the observation y can be adequately modeled with a selected few training instances. Clearly, this equation plays a role analogous to the LPM expression Eq. 6 mentioned previously. Note, however, that in LPM the notion of *sparsity* is, in effect, subsumed by that of *dimensionality reduction*.

Since the matrix H in EBCSR is built from the (possibly unbounded) set of observations available for training, the computation of the weight vector β may be prohibitive for large values of T. To circumvent this issue, existing implementations typically fall back to a nearest-neighbor paradigm, i.e, constructing $H_{d\times k}$ from k nearest samples (a form of quantization) to y in the acoustic feature space [13] and following the subsequent steps for classification.

Interestingly, LPM sidesteps the above unbounded problem by first performing a vector quantization step on the observations, so as to obtain a (bounded) set of N codebook entries ($N \ll T$). Then, the need for a sparseness constraint is effectively eliminated by converting the original acoustic description of training instances into a derived description involving unit-segment frequencies. In the same way that the EBCSR weight vector β weighs the “best” instances in the dictionary $H_{d\times k}$ to represent the data, LPM is able to represent each observation as a linear combination of the “best” data-driven speech units obtained after projection into the latent perceptual space.

Thus the two frameworks offer intriguing parallels. In particular, critical parameters are closely aligned: in LPM the size of the codebook N is analogous to k above, and the dimension of the LPM space R is analogous to d above.

4. EXPERIMENTAL SETUP

Classification experiments were conducted on the read speech TIMIT data using the standard NIST training and core test sets with known phoneme boundaries. A total of 140,225 and 7,215 phoneme instances were available for training and testing, respectively. The original 61 phoneme classes available in the corpus was reduced to 48 labels for training and the classification experiments were ultimately evaluated by mapping the 48 class labels to 39 labels. For feature vectors, 14-dimensional MFCC features with its delta and delta-delta components were extracted every 5 ms with a 10 ms Hamming window as acoustic features. The features were then subjected to linear discriminant analysis (LDA) using the reduced 48 phone labels.

The LPM codebook size N was set to 1000, the number of units W varied as $\{10^2, 10^3, 10^4\}$ and the dimensionality R varied in steps. For each parameter setting, we separately considered both short units (SU) consisting of 1-, 2- and 3-grams and long units (LU) obtained by dividing a phoneme segment in 30-40-30% sub-parts. Unit selection was performed as mentioned earlier using a combination of indexing power and empirical probability, as follows:
Method 1 (M01): This method focuses on units which appear only a handful of times in the training instances, so we used the indexing power Φ^i to rank all the units and only select the top W units for the subsequent LPM steps.
Method 2 (M02): The method focuses on units that possess high indexing power and high empirical probability both within a given phoneme class. The indexing power Φ^i is modified to Φ^i_p i.e, the indexing power within each phoneme class, and the final rank for each unit is obtained as:

$$\rho^i_p = \Phi^i_p * \pi^i_p \qquad (7)$$

after further weighting by the empirical probability of a unit within each class.

M01 prefers longer length units that sparingly occur in the collection while M02 balances both shorter units that occur frequently and longer length units. We can therefore expect M01 to perform better with long units (LU) derived by 30-40-30 partitioning of the phoneme instances, and M02 to perform better with short units (SU) up to 3-grams.

5. RESULTS

Phoneme classification performance for the proposed LPM approach with variable length units is shown in Figure 3. Short units (SU with 1-, 2- and 3-grams) are on the left and long units (LU with 30-40-30 partitioning) on the right, for different number of units W. Direct DTW-based classification using *complete* phoneme segments on quantized sequences results in an average performance of

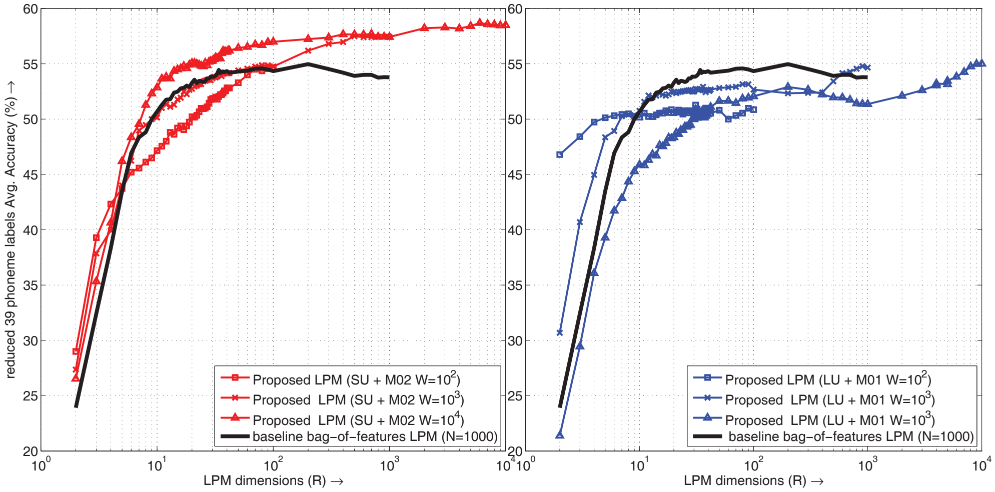

Fig. 3. Proposed LPM classification performance.

58.2%. The baseline LPM approach (shown in black) is the original LPM formulation [1] where only single-frame units (1-gram, bag-of-frames) are considered and the number of units W equals the codebook size $N = 10^3$. Its maximum performance is 55%. The proposed approach using variable length short and long units reveals interesting trends.

For the short units (SU), M02 ranking resulted in the best performance and as the number of units considered (W) increased, the classification performance also increased. For both $W = 10^3$ and $W = 10^4$, the performance exceeded the performance of the baseline LPM. Particularly for $W = 10^4$, the performance gain of about 1-4% can be observed for all R. For this case, the maximum performance is 58.4% which matches the direct DTW on complete phoneme segments in spite of using relatively low number of units $(W \ll M)$. For long units (LU), M01 ranking resulted in the best performance. Interestingly, the performance for very low LPM dimensions $(2 \leq \text{R} \leq 5)$ and small number of units $W = 100$, is up to 47% with about 1-2% improvement for larger dimensions. As the number of units increases, the performance for low dimensions decreases and the overall performance for larger LPM dimensions also decreases. For both $W = 10^3$ and $W = 10^4$, the maximum performance achieves the performance of the baseline LPM at higher dimensions.

6. CONCLUSION

Through a data-driven derivation of prototypical acoustic units in a latent space of low dimensionality, latent perceptual mapping allows for a viable template-like recognition strategy where models are closely aligned with perceived acoustic events. In this paper, we have extended the original bag-of-frames formulation to a more comprehensive bag-of-units framework which can take advantage of data-driven variable-length acoustic units. This in turn enables segment-level temporal information to be leveraged directly in LPM feature extraction. It also casts the framework as a generalization of other DTW-based methods that operate in the acoustic feature space. Experimental evidence shows that the ensuing acoustic modeling results in improved classification accuracy on the TIMIT task. Our results also show an interesting interplay between the length of the unit and the performance at a given dimensionality of the latent space. By appropriate choice of length of unit and number of units relatively high classification performance can be achieved in low dimensional latent space. These findings support our premise illustrated in Figure 2.

Future work will further refine our unit derivation approach. While we have presented a ranking-based approach to control the number of units, the clustering approach presented in [1] can also be adopted. As we already use dynamic time warping for the unit-document matrix appropriate temporal decoding procedures both during training and recognition can also be incorporated. Furthermore, we would also like to explore divergence measures for similarity computations [11] and further explore the parallels with sparse representations [13] discussed in Section 3.

7. REFERENCES

[1] S. Sundaram and J. Bellegarda, "Latent Perceptual Mapping: A New Acoustic Modeling Framework for Speech Recognition," *INTERSPEECH 2010, Makuhari, Japan.*, pp. 881–884, 2010.

[2] L. Deng and H. Strik, "Structure-Based and Template-Based Automatic Speech Recognition - Comparing parametric and non-parametric approaches," *Interspeech , Antwerp, Belgium*, 2007.

[3] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle, "Template-based Continuous Speech Recognition," *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 15, pp. 1377–1390, 2007.

[4] S. Axelrod and B. Maison, "Combination of Hidden Markov Models with Dynamic Time Warping For Speech Recognition," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 173–176, 2004.

[5] X. Sun and Y. Zhao, "Integrate Template Matching and Statistical Modeling for Speech Recognition," *INTERSPEECH 2010,*, pp. 173–176, 2010.

[6] G. Heigold, G. Zweig, X. Li, and P. Nguyen, "A flat direct model for speech recognition," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 3861–3864, 2009.

[7] A. Jansen and P. Niyogi, "Detection-based Speech Recognition with sparse point process models," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010.

[8] A. Mohamed and G. Hinton, "Phone Recognition using Restricted Boltzmann Machines," *IEEE International Conference on Acoustics Speech and Signal processing (ICASSP)*, 2010.

[9] M. De Wachter, K. Demuynck, D. Van Compernolle, and P. Wambacq, "Data Driven Example Based Continuous Speech Recognition," *In Proc. EUROSPEECH, Geneva, Switzerland*, 2003.

[10] S. Deerwester, S. T. Dumais, G. W. Furnas, T.K. Landaurer, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, vol. 6, no. 41, pp. 391–407, 1990.

[11] Hugo van Hamme, "HAC-models: A Novel Approach to Continuous Speech Recognition," *INTERSPEECH 2008, Brisbane, Australia*, pp. 2554–2557, 2008.

[12] J. Bellegarda, "Latent Semantic Mapping," *IEEE Signal Processing Magazine*, vol. 22, pp. 70–80, 2005.

[13] D. Kanevsky, T. N. Sainath, B. Ramabhadran, and D. Nahamoo, "Analysis of Sparseness and Regularization in Exemplar-based Methods for Speech Classification," *INTERSPEECH, Makuhari, Japan*, pp. 2842–2845, 2010.