PROGRESS IN EXAMPLE BASED AUTOMATIC SPEECH RECOGNITION

Kris Demuynck, Dino Seppi, Hugo Van hamme, Dirk Van Compernelle

Katholieke Universiteit Leuven, Department of Electrical Engineering - ESAT
Kasteelpark Arenberg 10, Bus 2441, B-3001 Leuven, Belgium
{Kris.Demuynck,Dino.Seppi,Hugo.Vanhamme,Dirk.VanCompernelle}@esat.kuleuven.be

ABSTRACT

In this paper we present a number of improvements that were recently made to the template based speech recognition system developed at ESAT. Combining these improvements resulted in a decrease in word error rate from 9.6% to 8.2% on the Nov92, 20k trigram, Wall Street Journal task. The improvements are along different lines. Apart from the time warping already applied within the DTW, it was found beneficial to apply additional length compensation on the template score. The single best score was replaced by a weighted k -NN average, while maintaining natural successor information as an ensemble cost. The local geometry of the acoustic space is now taken into account by assigning a diagonal covariance matrix to each input frame. Context sensitivity of short templates is increased by taking cross boundary scores into account for sorting the N best templates. Furthermore boundaries on the template segmentations may be relaxed. Finally context dependent word templates are now being used for short words. Several other variants that were not retained in the final system are discussed as well.

Index Terms— Speech Recognition, Template Based Recognition, Example Based Recognition, k Nearest Neighbours, DTW

1. INTRODUCTION

Speech recognition using an example based approach has gained significant attention over the past couple of years [1, 2, 3, 4, 5, 6]. At ESAT we have developed a large vocabulary template based speech recognition system. The original system was described in [1] with additional improvements reported on in [7, 8]. In this paper we present new significant enhancements over these previous systems, bringing it close in performance to state-of-the-art HMM systems.

The key idea behind our example based approach is: avoid the information loss resulting from abstracting large amounts of data into compact models such as HMMs. Instead, store the observed data enriched with some annotations and infer on-the-fly from the stored examples when dealing with a classification or recognition problem by finding those examples –chunks of audio with their corresponding annotations– that resemble the input data best. This on-the-fly inference has the potential advantage of making better use of the available data by focusing on the most relevant data, e.g. those examples of a phone which were uttered at the same speaking rate by a speaker with a similar accent, and originating from the same word (or syllable) as the word the recognizer hypothesizes.

The quality of the inference engine –typically k Nearest Neighbours (k -NN) based– is crucial to the performance of an example

Fig. 1. Word Graph with Phone Segment Annotations expanded to a Template Graph.

based system: it must be fast and yet at the same time make optimal use of the available data. One weakness in our existing framework is –given the single best Viterbi decoding strategy– the system’s inherent sensitivity to errors in the training database such as incorrect annotations, bad segmentations, highly unusual pronunciations, and inaccurate or missing phonetic transcriptions in the lexicon. Previous work such as data sharpening [7] and data pruning [9] aimed directly at mitigating these problems. Some of the techniques presented in this paper continue along this research line while other techniques aim to make better use of the available data.

The paper is organized as follows. First, we describe our baseline system. The following sections handle the major enhancements in the inference engine: (i) usage of weighted k -NN template scores instead of the single best Viterbi decoding, (ii) assigning a local sensitivity matrix (diagonal covariance) to each input frame, (iii) modifications to the dynamic time warping constraints and score computation, and (iv) usage of word based templates. The last sections deal with some open issues and possible further improvements.

2. BASELINE SYSTEM

The Wall Street Journal database was used for all experiments reported on in this paper. Results are given for the Nov92 20k open vocabulary non verbalized punctuation test set using the default trigram language model. The phonetic transcriptions for the train data and the 20k test lexicon were drawn from CMUdict 0.6d.

A baseline template based system was created according to the principles described in [8]: an HMM system generates word graphs enriched with phone segmentations; k -NN template lists are added to each phone; the word score is calculated with a Viterbi search through the templates after applying the template transition costs. This process is illustrated in figure 1. The baseline HMM system is trained on the WSJ0+1 SI-284 data comprising 81 hours from 284 speakers. SPRAAK [10] was used to create a conventional HMM system using Mel Spectra, incorporating vocal tract length normalization and mean subtraction, postprocessed by mutual information discriminant analysis (MIDA), as features. The HMM

This work was supported by FWO research project G.0260.07 “Telex”, FWO travel grant K.02.105.10N, the Sound-to-Sense EU Marie Curie Research Training Network (RTN-CT-2006-035561) and the John Hopkins 2011 Summer Workshop on Speech Recognition with Segmental CRFs.

system uses a shared pool of 32k Gaussians and 5875 cross-word context-dependent tied triphone states.

For the template system, all data from the 284 training speakers are used, without any cleaning for pronunciation or transcriptions errors. The baseline HMM system was used to segment the train database into 2826699 phone templates. The features used in the template system are the same as in the HMM system, except for additional data sharpening to reduce the influence of outliers [7]. Using the decision tree approach from [8], the phone templates were subdivided into 4219 cross-word context-dependent triphone classes, with a minimum of 256 templates per class. Context-dependent variants of the word arcs (see figure 1) are added so that the triphone class for each word boundary phone is uniquely defined by the word internal adjacent phone on the one side and all possible cross-word adjacent phones on the other side.

For development and parameter tuning we relied on the Dev92 development set. Optimization was done on either word error rate (WER) or on the product of the posterior probabilities on the correct path when WER was deemed too noisy.

3. WEIGHTED K -NN SCORES

Despite data sharpening, the single best Viterbi decoding approach used in [1, 7, 8] remains sensitive to outlier and mislabeling effects. One solution is to use the forward(-backward) algorithm in combination with a proper transformation of the scores instead of the Viterbi algorithm for decoding, similar to what was done in [11]. The built-in averaging of the scores in the forward algorithm automatically mitigates the impact of outliers. In this work, we opted for an even simpler solution: averaging the scores –again after a proper transformation– before the decoding, hence avoiding Viterbi/forward decoding at the template level altogether. Note that k nearest neighbour templates per phone arc (see figure 1) give rise to $k \times k$ transitions, resulting in a non-negligible cost for the template based Viterbi/forward decoding for typical values of $k = 50 \dots 100$.

The natural successor cost [1], a cost that is added on the transition between any two templates that are not immediate successors in the train database, has given consistent relative error reductions between 5 and 10%. Using weighted average scores requires another mechanism to preserve this valuable information. In the new implementation –before averaging– a natural successor cost is added to a template score if none of the adjacent template arcs is the natural successor. This is repeated for left and right context.

We also noticed that scores could best be adjusted based on the length of the phone arc under consideration. This is due to the fact that for shorter phone arcs there are (i) more candidate templates, (ii) less frames to match and hence less chance for badly matching frames, and (iii) the Itakura constraints [12] in the dynamic time warping (DTW) are less restrictive on the boundary frames.

Equation 1 shows the score length compensation and score averaging that was used in the final system, with l the length of the phone arc in frames, s_i the i^{th} best template score, c_i^l and c_i^r the left and right natural successor cost, $\alpha = 0.15$ and $N = 5$.

$$s' = \frac{1}{l^\alpha} \times \sqrt{\frac{N}{\sum_{i=1}^N \frac{1}{(s_i + c_i^l + c_i^r)^2}}} \quad (1)$$

This equation was chosen as it does a reasonable job in approximating the process of first transforming the template scores (norm₂ distances) into some probability measure, followed by the weighted average of the probabilities, followed by a back transformation from the average probability to an average template score.

Fig. 2. DTW Path Constraints.

4. LOCAL SENSITIVITY MATRIX

The investigation in distance measures for DTW in [13] lead to contradictory results: either the Euclidean or the Mahalanobis distance measure performed best depending on the task (phone classification versus continuous speech recognition). This behaviour could be related to whether one was looking in the near neighbourhood (single frame classification) of the input frame or whether one needed to also consider not so near reference frames, e.g. when matching longer templates.

By making the covariance (sensitivity) matrix needed for the Mahalanobis distance dependent on the input frame instead of the reference frame (more correctly, the triphone class the reference frame belongs to), most of the problems that adversely affect the Mahalanobis distance measure from [13] can be avoided. Since the Jacobian now depends on the input frame, it can be readily dropped. Furthermore, the properties of the distance measure now no longer change dramatically based on the (reference) frame one compares the input frame with. Yet, at the same time, the distance measure is adaptive to the local distribution (manifold) of speech data.

One available resource that already models the local distribution of the speech data is the shared pool of Gaussians from the HMM baseline system. We used the set of (diagonal covariance) Gaussians $\mathcal{N}(x; \mu_i, \Sigma_i)$, $i \in [1 \dots M]$ with corresponding a priori probabilities α_i as follows to assign a covariance Σ_x to an input frame x :

$$\Sigma_x = \frac{\sum_{i=1}^M \alpha_i \mathcal{N}(x; \mu_i, \Sigma_i)^\beta \Sigma_i}{\sum_{i=1}^M \alpha_i \mathcal{N}(x; \mu_i, \Sigma_i)^\beta} \quad (2)$$

By setting the parameter β to a value smaller than 1.0 (β equaled 0.4 in our experiments), Σ_x changes more slowly in function of x . From a theoretical point of view, β reflects the difference between the physical dimensionality of the observation vector x (39 in our experiments) and the intrinsic dimensionality of speech (which is somewhere between 4 and 11, depending on the phone [14]).

5. DTW SCORING

In this section we investigate the integration of the frame distances into a template distance by means of (dynamic) time warping.

Dynamic time warping aligns each input frame x_i to one reference frame y_j given some transition constraints so that the resulting sum of inter frame distances and transition costs is minimal. A common DTW implementation, which is also our reference implementation, allows the three types of transitions depicted in figure 2. Each of these transitions corresponds to a different rate of speech for the test data versus the reference template:

$T_0 : (x_{i-1}, y_j) \rightarrow (x_i, y_j)$ extreme slow rate of speech

$T_1 : (x_{i-1}, y_{j-1}) \rightarrow (x_i, y_j)$ same rate of speech

$T_2 : (x_{i-1}, y_{j-2}) \rightarrow (x_i, y_j)$ double rate of speech

Fig. 3. Dynamic Time Warping with Boundary Extensions.

The Itakura constraints [12] make the type T_0 and T_2 transitions more symmetric by allowing a T_0 transitions after a T_1 or T_2 transition only. This way, the minimal rate of speech of the test data is now at least half of that in the reference template (an alternation of T_0 and T_1 transitions). We also tried out a quite different warping strategy: linear warping. Given that the average length of a phone is only 8 frames, and given that the train database contains examples uttered at different speaking rates, dynamic warping may be inessential or even undesirable. Experiments with the different warping constraints were quite conclusive: the limitations imposed by the Itakura constraints improved the accuracy with respect to the baseline, while linear warping was clearly inferior to dynamic warping.

Another open issue with DTW is how to penalize non-diagonal (fast and slow speaking rate) transitions. This penalty can either be a fixed additive cost or a cost proportional to the inter frame distance (implemented by multiplying the local distance with some factor higher than 1.0). Whereas previous experiments [1] were inconclusive, the current WSJ setup showed a clear preference for additive costs. Additive costs also make more sense from a theoretical point of view since (i) additive costs can be interpreted as transition probabilities, similar to the transition probabilities in an HMM system, and (ii) the multiplicative costs will upscale the inter frame distances for those frames which need a non-diagonal warping; in other words more weight is assigned to the acoustic (dis-)similarity of those frames which show a non-standard behaviour.

Finally, we investigated the impact of promoting acoustic continuity when selecting the k -NN templates. This was done by taking the similarity between the audio to the left and right of a reference template and the audio surrounding a hypothesized phone or word arc into account during k -NN template selection. Figure 3 shows the most flexible configuration we investigated. DTW with backtracking over the extended region allows us to subdivide the DTW score into similarity measures for the central part and left and right surroundings. Characteristic for this approach is that it allows time warping in the surrounding regions which in turn allows for some automatic adjustment of the template boundaries. Relaxing the template boundaries has the potential benefit of making the system less dependent on the quality of the template boundaries and on the exact begin and end time of the phone and word arcs in the word graph.

We also experimented with a second setup which took the tem-

plate and phone/word arc boundaries as given. Since the left and right surrounding regions only comprise a few frames, time warping in these regions might be unnecessary, i.e. when calculating the acoustic similarity for the left and right l_x surrounding frames, a straightforward 1-to-1 alignment is assumed (the thick dashed alignment path in figure 3).

In both setups, the extended score (sum of the scores of the central, left and right surrounding regions) was only used for ranking the templates for k -NN selection. Decoding relied on the score of the central part only. Multiple experiments consistently showed a preference for quite large contextual windows (up to 9 frames wide) which may be surprising given that the context-dependent templates already have context sensitivity built in. Given these wider context windows it was in the end not surprising that the normal dynamic time warping should also be applied in these regions.

6. WORD BASED TEMPLATES

The natural successor cost (see [1] and section 3) improves the recognition accuracy by promoting the use of longer template sequences. A next logical step is to actually use longer templates. In this work, we experimented with whole word templates. Whole word templates are expected to be most effective for function words since these are typically not modeled well by generic phonetic transcriptions and there are sufficient examples available to be used as reference templates by themselves.

The segmentation of the train database into word templates was done with the baseline HMM system, assuring consistency between the phone and word segmentations. The word templates are handled identically to the phone templates, using the same parameters for length compensation and averaging (section 3), local sensitivity distance measure (section 4) and DTW scoring (section 5). Similar to the phone templates, context dependent variants were generated when sufficient examples were available. The questions for the decision tree for a specific word were limited to those applied to the left context of the word's initial phone and to the right context of the word's final phone. This assures that the word context dependency is a subset of the phone context dependency, which simplifies the decoding process.

We provided for a smooth transition between word and phone modeling for those words with no or only a small number of templates (e.g. non-function words). In those cases the word score was replaced by an average of the word score w_s and the phone score w_p , weighted with the number of context dependent word examples N_w that were returned by the k -NN search:

$$s'_w = s_w \times \frac{N_w}{k} + s_p \times \left(1 - \frac{N_w}{k}\right), \quad (3)$$

with k the maximum number of near neighbour templates the inference engine looked for, which equaled 75 in our experiments.

7. RESULTS AND DISCUSSION

Table 1 summarizes the main results. Hereunder, we provide additional information and insight and give our main conclusions.

Score averaging & adjustment: The boost in performance thanks to the weighted average template scores was significant mainly because we managed to keep exploiting the natural successor concept. Natural successor costs are no longer applied on individual inter template transitions but are now incorporated in the single ensemble DTW score as an average natural successor cost for the selected k -NN neighbours.

System	Dev92	Nov92
initial template system		9.6%
+ score averaging & adjustment	10.6%	8.9%
+ local sensitivity & boundary extension	10.3%	8.5%
+ word templates	10.0%	8.2%

Table 1. Baseline system and impact of the successive improvements. WER on the WSJ Dev92 and Nov92 20k trigram task.

Local sensitivity & distance metric: For understanding the impact of the local distance modeling, it is important to realize that our raw features have undergone a MIDA transformation. MIDA is not only discriminative, but also decorrelates the features and does a proper scaling of the axes. The outcome is a feature space where an Euclidean distance metric performs remarkably well. Hence the modeling of the local manifold, which is not discriminative in its approach, provides only incremental improvements over a single global MIDA transform. When using for example Mel-Cepstra as features, much larger improvements can be observed.

Instead of an L_2 based distance measures, one may also opt for radial different metrics. One promising alternative is the sparse decompositions proposed by [6].

Boundary extension: Context-dependent templates enforce continuity at the symbolic (phone) level. Natural successors promote acoustic continuity, although in a limited and strict sense only. The consistent positive results obtained with both methods indicate that it is important to foster symbolic and acoustic template continuity. The “boundary extension” does exactly that. While it tackles problems due to poor segmentations in the database, its main contribution is that, by using overlapping segments, the k -NN lists of adjacent phones contain more natural successors (from 5.6% for $l_x = 0$ to 12.7% for $l_x = 9$) and hence favor a greater naturalness between adjacent template ensembles.

Multi-phone templates: Word templates are well suited for the frequent (function) words. Less frequent words may benefit from smaller more generic units, e.g. syllables. The current framework already allows for multiple levels of symbolic representation, so word and syllable units can even be combined.

Computational load: The use of ensemble scores (k -NN weighted average template scores) reduces the decoder complexity significantly: the single best template decoder had a search space that was an order of magnitude larger. By means of a careful implementation, the overhead of the template based score computation was reduced to a 0.1xRT load on a 12-core machine. Hence the current implementation lends itself much better than its predecessors for being used with larger databases. The template induced overhead can be kept constant by creating more context dependent variants and by preselecting templates on the basis of their length (within a small range of the test segment length) when handling large databases.

8. CONCLUSIONS

In this paper we presented a set of relatively simple techniques for improving template based speech recognition, leading to a 15% relative word error rate reduction. One key advance was the migration from a single best template decoder to a phone/word decoder which uses ensemble scores (k -NN weighted average template scores). Furthermore, by maintaining the natural successor costs—now calculated based on adjacent template ensembles—and by adding “boundary extensions”, template continuity was improved. While already close to state-of-the-art HMM systems, the template based approach has still room for significant improvements by

exploiting the wealth of meta-information that is present in the template lists. Such an approach is presented in a companion paper [15]. Further improvements may come from newly emerging techniques such as sparse decompositions [6].

9. REFERENCES

- [1] Mathias De Wachter, Mike Matton, Kris Demuynck, Patrick Wambacq, Ronald Cools, and Dirk Van Compernelle, “Template based continuous speech recognition,” *IEEE Trans. on ASLP*, vol. 15, pp. 1377–1390, May 2007.
- [2] Viktoria Maier and Roger K. Moore, “Temporal episodic memory model: An evolution of Minerva2,” in *Proc. INTERSPEECH*, Aug. 2007, pp. 866–869.
- [3] V. Ramasubramanian, Kaustubh Kulkarni, and Bernhard Kämmerer, “Acoustic modeling by phoneme templates and modified one-pass DP decoding for continuous speech recognition,” in *Proc. ICASSP*, Apr. 2008, pp. 4105–4108.
- [4] Xie Sun and Yunxin Zhao, “Integrate template matching and statistical modeling for speech recognition,” in *Proc. INTERSPEECH*, Sept. 2010, pp. 74–77.
- [5] Ladan Golipour and Douglas O’Shaughnessy, “Phoneme classification and lattice rescoring based on a k -NN approach,” in *Proc. INTERSPEECH*, Sept. 2010, pp. 1954–1957.
- [6] Dimitri Kanevsky, Tara N. Sainath, Bhuvana Ramabhadran, and David Nahamoo, “An analysis of sparseness and regularization in exemplar-based methods for speech classification,” in *Proc. INTERSPEECH*, Sept. 2010, pp. 2842–2845.
- [7] Mathias De Wachter, Kris Demuynck, and Dirk Van Compernelle, “Outlier correction for local distance measures in example based speech recognition,” in *Proc. ICASSP*, Apr. 2007, vol. IV, pp. 433–436.
- [8] Sébastien Demange and Dirk Van Compernelle, “HEAR: an hybrid episodic-abstract speech recognizer,” in *Proc. INTERSPEECH*, Sept. 2009, pp. 3067–3070.
- [9] Dino Seppi and Dirk Van Compernelle, “Data pruning for template-based automatic speech recognition,” in *Proc. INTERSPEECH*, Sept. 2010, pp. 901–904.
- [10] Kris Demuynck, Jan Roelens, Dirk Van Compernelle, and Patrick Wambacq, “SPRAAK: An open source speech recognition and automatic annotation kit,” in *Proc. INTERSPEECH*, Sept. 2008, p. 495.
- [11] Kris Demuynck, Dirk Van Compernelle, and Patrick Wambacq, “Doing away with the Viterbi approximation,” in *Proc. ICASSP*, May 2002, vol. I, pp. 717–720.
- [12] Fumitada Itakura, “Minimum prediction residual principle applied to speech recognition,” *IEEE Trans. on ASSP*, vol. 23, no. 1, pp. 67–72, Feb. 1975.
- [13] Mathias De Wachter, Kris Demuynck, Patrick Wambacq, and Dirk Van Compernelle, “Evaluating acoustic distance measures for template based recognition,” in *Proc. INTERSPEECH*, Aug. 2007, pp. 874–877.
- [14] Mattias Nilsson and Bastiaan Kleijn, “On the estimation of differential entropy from data located on embedded manifolds,” *IEEE Trans. on Information Theory*, vol. 53, no. 7, July 2007.
- [15] Kris Demuynck, Dino Seppi, Dirk Van Compernelle, Geoffrey Zweig, and Patrick Nguyen, “Integrating meta-information into exemplar-based speech recognition with segmental conditional random fields,” in *Proc. ICASSP*, 2011, submitted.