INTEGRATING META-INFORMATION INTO EXEMPLAR-BASED SPEECH RECOGNITION WITH SEGMENTAL CONDITIONAL RANDOM FIELDS

*Kris Demuynck, Dino Seppi, Dirk Van Compernolle**

Katholieke Universiteit Leuven - ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

Patrick Nguyen, Geoffrey Zweig

Microsoft Research
Redmond, WA

ABSTRACT

Exemplar based recognition systems are characterized by the fact that, instead of abstracting large amounts of data into compact models, they store the observed data enriched with some annotations and infer on-the-fly from the data by finding those exemplars that resemble the input speech best. One advantage of exemplar based systems is that next to deriving what the current phone or word is, one can easily derive a wealth of meta-information concerning the chunk of audio under investigation. In this work we harvest meta-information from the set of best matching exemplars, that is thought to be relevant for the recognition such as word boundary predictions and speaker entropy. Integrating this meta-information into the recognition framework using segmental conditional random fields, reduced the WER of the exemplar based system on the WSJ Nov92 20k task from 8.2% to 7.6%. Adding the HMM-score and multiple HMM phone detectors as features further reduced the error rate to 6.6%.

***Index Terms*—** Speech Recognition, Template Based Recognition, Example Based Recognition, k Nearest Neighbours, Conditional Random Fields, SCARF

1. INTRODUCTION

Exemplar (also referred to as template or episodic) based approaches to automatic speech recognition have recently aroused a revival of interest [1, 2, 3, 4, 5, 6]. Exemplar based approaches are characterized by the fact that, instead of abstracting large amounts of data into compact models, they store the observed data enriched with some annotations and infer on-the-fly from the data by finding those exemplars that resemble the input speech best. One of the benefits of example based approaches is that they avoid the information loss resulting from abstracting data into compact models, i.e. they preserve all details in the training data such as trajectories, temporal structure and (fine) acoustic details.

Another advantage of exemplar based systems, one that is typically not exploited, is that next to deriving a goodness of fit score for the hypothesized phone or word, one can easily derive a wealth of meta-information concerning the chunk of audio under investigation. Examples are: who is the speaker, what is his/her gender and age, what dialect does he/she speaks, what is the speaking rate, where are the word boundaries and/or sentence boundaries, and what is the underlying prosodic structure. Note that the metadata features can pertain not just to one template, but to the ensemble of matching templates. An example thereof is speaker entropy. Statistics on the metadata can be obtained from the same k nearest neighbour (k-NN) search as used for deriving the primary recognition score.

Collecting such meta-information and integrating it into the overall recognition framework is the main topic of this paper. Since the number of relevant meta-statistics (features) that can be derived is fairly large, an efficient scheme is needed to rank the features and to combine these features with the primary recognition score. We found the segmental conditional random fields (SCRF) framework [7], and more in particular the SCARF toolkit [8] to be a powerful and fast method to perform such complex optimizations.

The remainder of this paper is organized as follows. First, we describe our baseline HMM and template setup. Next we present the meta-information that is extracted from k-NN template lists and introduce the SCRF approach used to integrate the extra features into the setup. Finally we present the results of the template system in isolation and in combination with additional HMM systems.

2. BASELINE SYSTEM

The Wall Street Journal (WSJ) database was used for training, developing and testing the template system and the additional HMMs for system combination. Results are presented on the Nov92 20k open vocabulary test set using the default trigram LM. Training is done on the SI-284 data from WSJ0+1 comprising 81 hours from 284 speakers. The phonetic transcriptions for the training data and the 20k test lexicon were drawn from CMUdict 0.6d. The SCARF toolkit was used to find the optimal values for all tunable parameters. For the template system, we relied on the non verbalized punctuation part of the Dev92 development set for all development and parameter tuning. For the system combination with the HMM systems, the training data were used in a leaving one-speaker-out configuration.

2.1. Baseline HMM system

SPRAAK [9] was used to create a conventional HMM system using Mel Spectra, incorporating vocal tract length normalization and mean subtraction, and postprocesed by mutual information discriminant analysis (MIDA), as acoustic features. This baseline HMM system uses a shared pool of 32k Gaussians and 5875 cross-word context-dependent tied triphone states.

2.2. Baseline template system

The work on template features started from a system operating according to the principles described in [10], i.e. the baseline HMM system generates word graphs enriched with phone segmentations after which each word arc score is replaced with the sum of the corresponding context-dependent phone template scores. The template scores are calculated by means of dynamic time warping (DTW).

*This work was supported by FWO research project G.0260.07 "Telex", FWO travel grant K.2.105.10N, 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

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 phone templates. The acoustic features adopted in the template system are the same as in the HMM system, except for additional data sharpening to reduce the influence of outliers [11]. The decision tree approach from [10] was used to sub-divide the phone templates in 4219 cross-word context-dependent triphone classes. New context-dependent variants of the word arcs are introduced in the word graphs to accommodate the template triphone classes.

A first set of improvements on this initial template system refines the way the acoustic similarity score is calculated. These improvements are described in more detail in a companion paper [12]. Hereunder, we briefly list the techniques used along with those details that are relevant to the work reported on in this paper.

Score averaging: The setup in [10] searches for the single best template sequence for a given input signal by means of Viterbi decoding. This renders the system sensitive to errors in the training data such as incorrect annotations, bad segmentations, highly unusual pronunciations, and inaccurate phonetic transcription in the lexicon. Averaging the template scores before decoding mitigates this problem and simplifies the search space for the SCRF decoding tremendously.

Score adjustments: (Non-)Natural successor costs are added to the DTW scores whenever some piece of input speech can only be explained by mixing and matching templates which did not form a sequence in the original train database. The "natural successor cost" promotes the use of longer stretches of reference data to explain the incoming signal [1]. The DTW score is also adjusted in function of the duration of the phone arc under consideration [12]. This adjustment compensates for 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 [13] in the DTW are less restrictive on the boundary frames.

Local sensitivity (data covariance): In HMM systems, the covariances in the Gaussians allow for a more precise modelling of the local distribution (manifold) of speech data. A previous attempt to introduce covariances in a template system showed mixed results [14]: since the covariances depended on the class the references frame belongs to, the comparison of distances between a single input frame (of unknown class) with reference frames belonging to different classes became tricky as the properties of the distance measure could vary greatly depending on the reference class (and hence covariance). Linking the covariance to the input frame instead of to the reference frame avoids these problems. The covariance Σ_x in the neighbourhood of an input frame x is estimated by means of equation 1, with $\mathcal{N}(x;\mu_i,\Sigma_i)$ the pool of diagonal covariance Gaussians from the baseline HMM, α_i the corresponding a priori probabilities, and β a tunable parameter.

$$\Sigma_x \;=\; \frac{\sum_i \alpha_i \mathcal{N}(x;\mu_i,\Sigma_i)^\beta \Sigma_i}{\sum_i \alpha_i \mathcal{N}(x;\mu_i,\Sigma_i)^\beta} \qquad (1)$$

Acoustic boundary extensions: The use of context-dependent phone templates [10] biases the k-NN template search to those templates that are expected to fit well with the surrounding phones. Instead of relying only on the symbolic phone labels, one could also compare the signal surrounding the phone arc and template and require a good acoustic match. An added benefit of this scheme is that the system becomes less dependent on the quality of the template boundaries and on the exact begin and end time of the phone/word arcs in the word graph. For the experiments reported on in the paper, the time intervals spanned by the phone arc (input signal) and reference templates were expanded with $l_x = 9$ frames to the left and right before doing the DTW. The "extended" DTW score was used for selecting the k best matching templates. For the actual decoding, the score was limited to that part of the DTW score corresponding to the time interval spanned by the phone arc.

System	WER
HMM baseline	7.3%
initial template system	9.6%
+ score averaging & adjustment	8.9%
+ local sensitivity & boundary extension	8.5%
+ word templates	8.2%

Table 1. WER of the baseline systems (WSJ Nov92 20k trigram).

Word templates: Next to promoting the use of longer template sequences by means of the natural successor cost, we also improved the system by adding whole word templates to the inventory of acoustic units for template matching.

2.3. Results

Table 1 gives the word error rate (WER) of the baseline HMM system and the template system with and without the improvements listed above. As can be seen, a sequence of relatively simple techniques bridges 60% of the performance gap between the initial template system and the state-of-the-art baseline HMM system. This is a promising trend, especially since large vocabulary template based recognition is still in its early stages of development while HMM based recognition is a well established approach.

3. META INFORMATION DERIVED FROM K-NN LISTS

When dealing with a classification or recognition problem, example based systems infer the necessary statistics on-the-fly from the stored examples by finding those examples –chunks of audio with their corresponding annotations– that resemble the input data best. The techniques listed in section 2.2 aim directly at improving the primary statistic, namely the goodness of fit score for the hypothesized words. In this section we will focus on what other statistics can be derived from the k-NN template lists.

The starting point for deriving these secondary statistics is the k-NN list of phone templates. Each template in the list is characterized by its DTW score and the meta-information (annotations) associated with that template. For this work, the following meta-information was considered: (i) the template index (natural order of the templates in the train database), (ii) the phone label, (iii) the phone duration, (iv) the speaker ID, (v) the word the phone originated from, and (vi) the position of the phone in the word (word initial, final, ...)

Instead of just counting occurrences in the k-NN list, we opted to weigh each occurrence with the template's probability measured by converting the DTW score by means of equation 2

$$w_i = \exp\left(-0.5\frac{d_i^{(c)} + d_i^{(l)} + d_i^{(r)}}{l_a + 2l_x}\right),\; p_i = \frac{w_i}{\sum_{j=1}^{k} w_j}, \qquad (2)$$

with l_a the number of frames spanned by the phone arc, $l_x = 9$ the number of extra surrounding frames used in the DTW-alignment, $d_i^{(c)}$ and $d_i^{(l/r)}$ the DTW score corresponding to template i for the l_a central and l_x surrounding left/right frames respectively.

In view of the log-linear feature combination scheme employed in the SCARF toolkit, the secondary statistics were, when possible, converted to values behaving like a log probability normalized with their expected values under the condition that the measured property is true. The normalization compensates for the a priori distribution of the meta-information. Furthermore, given that the values of the primary features such as HMM and DTW scores are proportional to the number of frames, the secondary statistics were, when found to be beneficial, multiplied with the number of frames in the word. The combination of the statistics calculated on the phone arcs to word level measures was done with a weighted sum, the weights being proportional to the duration of the phones.

Word Position: A first set of features measures the consistency between the hypothesized word boundaries and word boundary statistics derived from the audio signal by means of harvesting the meta-information associated with the k-NN template lists. Given a chunk of audio corresponding to a hypothesized context-dependent phone arc and the corresponding k-NN template list, the feature f^{wi} measuring the property of being word initial is calculated as follows:

$$c_1 = \sum_{i=1}^{k} p_i I_{\text{wi}}(\text{template}_i) \quad (3)$$

$$C_1 = \text{cnt}(\text{wi}, \text{cd-phone}) \quad (4)$$

$$C_0 = \text{cnt}(\neg\text{wi}, \text{cd-phone}) \quad (5)$$

$$C_1' = \frac{C_1 K}{C_1 + K} \quad (6)$$

$$C_0' = \min\left(\frac{C_0 K}{C_0 + K}, K - C_1'\right) \quad (7)$$

$$f^{\text{wi}} = \log\left(\frac{c_1 + \epsilon}{\frac{C_1'}{C_1' + C_0'} + \epsilon}\right) \quad (8)$$

c_1 is the weighted average of the indicator function $I_{\text{wi}}(\cdot)$ measuring whether the template template$_i$ is a word initial template in the train database. $C_{1/0}$ are the counts over all train templates (a priori distribution) of word-initial and non-word-initial occurrences for a given context-dependent phone. The ratio $C_1'/(C_1' + C_0')$ is an ad-hoc estimate for the expected average value of c_1 when the property "word initial" is true. Note that the expected value converges to 1.0 if there are enough positive examples ($C_1 \gg K$). The mapping from $C_{1/0}$ to $C_{1/0}'$ mimics the effect of looking only at the k nearest templates and weighting their importance ($K \leq k$). K, k and ϵ were set to 50, 75 and 0.1 respectively.

For each word, three word position features were derived, namely f^{wi} which measures the property of being word initial for the word initial phone arc, f^{wf} which measures the property of being word final for the word final phone arc, and f^{wm} which averages the property of being word internal over the remaining phone arcs.

Word Identity: The "word identity" feature measures the consistency between the hypothesized word ID and the word the phone templates were drawn from. This feature is calculated in the same way as the "word position" features, using proper indicator functions and counts.

Speaker Entropy: For each phone arc, the speaker entropy over the k-NN template list was calculated. Equation 2 was used to assign a probability to each template and hence each speaker. The rationale behind this feature is that phone realizations which are supported by a wide variety of speakers are more reliable than those for which only templates of a single speaker were selected.

Warping Factor: This feature compares the average duration of the templates in the k-NN list with the duration of the phone arc by means of equation 9, with l_i the duration of template i and l_a the duration of the phone arc.

$$f^{\text{warp}} = -\left|\log\left(\frac{\sum_{i=1}^{k} p_i l_i}{l_a}\right)\right| \quad (9)$$

The value of f^{warp} will be close to zero when a phone realization is supported by a set of templates with durations that are nicely spread around the duration of the phone arc and will be negative otherwise. The feature is thought to be useful for penalizing hypothesized phone realization with aberrant durations or to adjust the DTW score if the non-diagonal penalty cost is set too high or too low.

Natural Successors: Given a phone arc and the k-NN template list, the "natural successor" concept sets forth that a template with index m is more credible if template $m-1$ was selected for the phone arc to the left and/or if template $m+1$ was selected for the phone arc to the right. In other words, extra backing is given to a hypothesized sequence of phone arcs if the sequence of phone arcs matches well to templates that form a continuous sequence in the train database.

For each k-NN template list, we measure the (probabilistic) fraction v_a of templates for which (i) the natural predecessor could be found to the left, (ii) the natural successor could be found to the right, and (iii) both the natural predecessor and successor could be found. The three final word level features are formed by summing $l_a \times \log(v_a + \epsilon)$ over all phones arcs a in the word.

4. OPTIMIZATION AND INTEGRATION WITH SCARF

We used segmental conditional random fields [7], and more in particular the SCARF toolkit [8] to incorporate the different feature values in a single discriminative speech recognition framework. When operating on word graphs and with word level features, the SCARF training repeats the following steps:

1. compose the word graph with the language model (LM) finite state transducer;
2. calculate the word log likelihoods as a linear combination of feature values for that word (arc in the composed graph), the logarithm of the LM probability being one of the features;
3. calculate the word posterior probabilities by means of the forward-backward algorithm;
4. adjust the weights for the linear combination of feature values so that the product of word posteriors for the correct word sequence increases.

SCARF was used to perform the joint optimization of all weights in the linear feature combination. This includes the weights for the primary phone-based DTW-score, the word-based DTW-score, the log LM probabilities and all features relying on meta-information. A grid search based on the product of the posterior probabilities on the correct path as returned by SCARF, was used to optimize the system internal parameters such as the value of β in equation 1 or the non-diagonal cost in the DTW algorithm.

5. RESULTS

5.1. Template Based System with Meta Information

Table 2 shows the impact of adding meta-information based features to the baseline template system. Each (set of) features decreases the WER between 2% and 4% relative. Combining all features,

Extra features	Dev92	Nov92
/	10.0%	8.2%
Word Position	9.7%	8.0%
Word Identity	9.8%	7.9%
Speaker Entropy	9.8%	8.0%
Warping Factor	9.6%	7.9%
Natural Successors	9.7%	8.0%
All	9.3%	7.6%
+ baseline HMM	8.6%	6.8%
+ phone detectors	/	6.6%

Table 2. WER on the development and on the evaluation test set when adding extra features to the baseline template system.

brings the performance of the template based system close to that of the HMM system. The resulting 7% relative WER decrease indicates that the information provided by the different meta-information based features is largely complementary.

5.2. System Combination with HMMs

The SCARF toolkit makes it easy to add extra features to an existing system. One readily available feature is the HMM based word score. Combining the HMM and the template system in this way reduced the WER on the development and test set to 8.6% and 6.8% respectively. The large relative improvements indicate that the template and HMM system have different strengths and weaknesses. This is surprising since the template system is basically built on top of the HMM system, re-using for example the pre-processing, the Gaussian pool, and the train database segmentation.

The SCARF toolkit not only allows word level scores to be combined, but also promotes discrete (sub-word) detectors such as the single best phone sequence. Detector events are automatically converted to word level scores, either by means of a Levenshtein distance or by means of automatically detected (word,phone-sequence) relations [7, 8]. We added four phone detector streams to the setup. The first (primary) phone detector consisted of the baseline HMM combined with a bigram phone LM estimated on the train database. Three additional phone detectors were derived from variations on the baseline HMM system with different pre-processings, decision trees and sizes of the Gaussian pool. The best results were obtained with automatically detected (word,phone-pair) relations on the primary phone detector stream combined with Levenshtein features for all four phone detector streams. The SCARF training was run over the train database in order to provide enough training material for learning the (word,phone-pair) relations. A leaving-one-speaker-out approach was used to allow the re-use of the training data as development data for SCARF. When combining the phone detectors with the HMM baseline, the WER drops to 6.8%. Note that it took three complementary HMM systems to get the same performance boost as observed when combining with a single template based system. Adding the template system lowers the WER to 6.6%.

6. CONCLUSIONS

In this paper we showed that in exemplar based recognition systems, additional features based on meta-information such as the speaker ID can be readily derived. These features can be efficiently integrated in a template based recognition framework using segmental conditional random fields. More in particular, the SCARF toolkit allowed the heterogeneous knowledge sources to be integrated in a discriminative framework with little effort. Adding the meta-information based features lowered the WER of the template based system by 7% relative. We also showed, by means of system combination, that the template and HMM systems have different strengths and weaknesses, and that achieving the same improvement by means of combining different HMM systems requires more effort. Combining all systems resulted in a 6.6% WER, a 9.6% relative improvement over a monolithic state-of-the-art HMM system.

7. REFERENCES

[1] Mathias De Wachter, Mike Matton, Kris Demuynck, Patrick Wambacq, Ronald Cools, and Dirk Van Compernolle, "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] Geoffrey Zweig and Patrick Nguyen, "A segmental CRF approach to large vocabulary continuous speech recognition," in *Proc. ASRU*, Dec. 2009, pp. 152–157.

[8] Geoffrey Zweig and Patrick Nguyen, "SCARF: A segmental conditional random field toolkit for speech recognition," in *Proc. INTERSPEECH*, Sept. 2010, pp. 2858–2861.

[9] Kris Demuynck, Jan Roelens, Dirk Van Compernolle, and Patrick Wambacq, "SPRAAK: An open source speech recognition and automatic annotation kit," in *Proc. INTERSPEECH*, Sept. 2008, p. 495.

[10] Sébastien Demange and Dirk Van Compernolle, "HEAR: an hybrid episodic-abstract speech recognizer," in *Proc. INTERSPEECH*, Sept. 2009, pp. 3067–3070.

[11] Mathias De Wachter, Kris Demuynck, and Dirk Van Compernolle, "Outlier correction for local distance measures in example based speech recognition," in *Proc. ICASSP*, Apr. 2007, vol. IV, pp. 433–436.

[12] Kris Demuynck, Dino Seppi, Hugo Van hamme, and Dirk Van Compernolle, "Progress in example based automatic speech recognition," in *Proc. ICASSP*, May 2011, submitted.

[13] Fumitada Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. on ASSP*, vol. 23, no. 1, pp. 67–72, Feb. 1975.

[14] Mathias De Wachter, Kris Demuynck, Patrick Wambacq, and Dirk Van Compernolle, "Evaluating acoustic distance measures for template based recognition," in *Proc. INTERSPEECH*, Aug. 2007, pp. 874–877.