Boosting HMM Performance with a Memory Upgrade

Mathias De Wachter, Kris Demuynck and Dirk Van Compernolle

Katholieke Universiteit Leuven –Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven
{mathias.dewachter,kris.demuynck,dirk.vancompernolle}@esat.kuleuven.be

Abstract

The state-of-the-art in automatic speech recognition is distinctly Markovian. The ubiquitous 'beads-on-a-string' approach, where sentences are explained as a sequence of words, words as a sequence of phones and phones as a sequence of acoustically stable states, is bound to lose a lot of dynamic information. In this paper we show that a combination with example-based recognition can be used to recapture some of that information. A new approach to combine Hidden Markov Model (HMM) and phone-example-based continuous speech recognition is presented. Experiments show that the combination outperforms the HMM recognizer, and indicate that adding long-span information is especially beneficial.
Index Terms: example-based speech recognition, episodic, DTW.

1. Introduction

State-of-the-art ASR systems have had HMMs under the hood for over a quarter of a century now. Virtually all successes in the field owe to the great flexibility and scalability of the HMM modeling framework. Still, almost everybody agrees that this surefire framework has theoretical weaknesses. The unrealistic underlying independence assumptions and the general 'beads-on-a-string' approach where speech is explained as a sequence of words, words as a sequence of phones and phones as a sequence of states, has been drawing criticism from all comers. But pointing out its weaknesses has proved far easier than coming up with better alternatives. Some extensions such as segmental HMMs or 'super-models' such as DBNs have generally been able to show superior performance on small or well-chosen tasks, but overall, HMMs have retained their status [1, 2].

Over the past few years we have investigated pure example-based continuous speech recognition with promising results [3, 4, 5]. Example-based (or *template*-based) recognition captures more segmental information than HMMs on two levels: dynamic time warping matches acoustic paths rather than simplifying to three 'stable' states, and a mechanism that uses extra (non-verbal) information about the templates ensures smooth paths (see Section 2.4). The smoothness of these paths is in fact long-span information that cannot easily be used in the HMM framework. A major downside of example-based modeling lies in the acoustic distance calculation at the frame level. Mainstream DTW uses simple Euclidean distances in some acoustic space, while HMMs have state-dependent covariance matrices, i.e. state-dependent distance measures. In previous papers we showed that porting these state-dependent distance measures to the example-based framework results in large performance gains [4, 5].

However, while we were able to show that example-based recognition can outperform context-independent HMMs, state-of-the-art context-dependent HMMs have so far remained out of reach. Still, given the differences between the HMM and DTW approach, it is not unreasonable to assume that the scores are complementary and can be fruitfully combined. Combining scores in small word-based systems has shown improvements of up to 20% [6, 7]. However, the extension to a large vocabulary phone-based system may not be trivial due to the shear size of the DTW search space and the required optimizations.

2. Combined recognizer architecture

2.1. Concept

When combining HMMs and example-based recognition, an important issue is *at which stage* to combine hypotheses. Since HMMs use phone models to form hypotheses, the phone-level is the first opportunity for combination, while combining complete sentence hypotheses is the other extreme.

The latter is represented by the conceptually simplest approach, where an N-best list provided by the HMM recognizer is rescored with the DTW recognizer. This approach was successful for digit-string recognition in [7]. Rescoring whole sentences with a phone-example-based DTW recognizer is impractical however: finding the best sequence of phone templates with the corresponding time alignment is a challenging task given the shear amount of phone templates even moderately sized training databases contain. A possible solution would be the use of bottom-up template selection techniques [3]. Rescoring an HMM word graph would also need bottom-up selection, as especially for long words the number of template hypotheses based on different within-word time alignments is still too large for an exhaustive search.

If the combination is done at the phone level, however, the use of phone boundary timing information allows a complete database search, ensuring that the best matching templates are used. Phone-based combination fits well with a 2-layer decoder [8]. In the first layer the HMM system creates a dense phone graph. Any subsequent decoding (pure HMM, pure DTW or HMM and DTW combined) then starts from this phone network. The graph's phone boundary timings can be used either as absolute givens or to limit DTW phone template transitions to a small window. Preliminary experiments showed that allowing a small time window for phone transitions did not provide better results, so all experiments reported on in this paper use the 'strict' boundary policy.

Figure 1 shows a block diagram of the combined system. The upper part consist of the HMM phone decoder, resulting in the HMM-based phone graph. The lower part uses template matching to produce the template graph. Template concatenation costs are added when template scores are added to a running hypothesis. These costs are the reason that it is not sufficient to simply use the

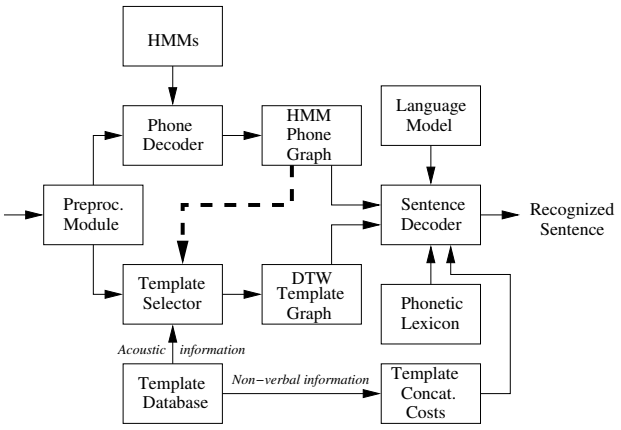

Figure 1: Block diagram of the combined system.

score of the acoustically best matching template in the database to rescore the corresponding arc in the HMM phone graph. The sentence decoder combines the HMM and template graphs with the concatenation costs, language model and lexicon to find the best recognition. The dashed arc indicates that the DTW template selection is constrained by the HMM phone graph.

The following sections discuss the system in detail.

2.2. HMM phone decoder

The ESAT/PSI speech group has a fully in-house developed state-of-the-art speech recognition system [9, 10]. For the RM model we used our default shared Gaussian approach, i.e. the density function for each of the 791 cross-word context-dependent tied states is modeled as a mixture over an arbitrary subset of Gaussians drawn from a global pool of 7487 Gaussians. The mixtures use on average 104.7 Gaussians to model the 36 dimensional observation vector. The 36 dimensions were obtained by means of a mutual information based discriminant linear transformation (mida) on 24 MEL spectra and their first and second order time derivatives [11].

A generic bottom-up phone decoder (as suggested in [8]) uses no lexicon or language model information. In this work we decided to constrain the phone decoder in the first layer to only these phone sequences that are allowed by the lexicon and the language model. An example phone graph for a short sentence is shown in figure 2. The nodes contain timestamps. The arcs contain the phone identity and the corresponding HMM score. Being based on context dependent phones, the graph also encodes the context constraints used in the HMM system. Hence, several nodes with the same timestamp exist. Furthermore, our graph construction algorithm automatically removes all sub-optimal transition boundaries between any two phones (given the context constraints imposed by the HMM system).

2.3. Template selector

The DTW system's preprocessing also uses the mida transform based on the same raw features, but here only the 25 most informative dimensions are retained. The frame-level distances are scaled locally, as described in [4]. Contrary to the HMM graph, context dependency constraints are not encoded in the template graph. Instead, the template number is added on each arc, allowing the on-line calculation of context costs. For each unique triplet (start frame, end frame, phone id) in the HMM graph, the n best templates –n is specified in Section 3– are added to the template graph.

2.4. Template concatenation

The sentence decoder adds concatenation costs based on non-verbal information about gender and original acoustic context of the templates [3]. This corresponds to multiplying with a prior template string probability in the following model for example-based speech recognition:

$$(\hat{\mathbf{W}}, \hat{\mathbf{T}}) = \underset{\mathbf{W},\mathbf{T}}{\operatorname{argmax}}\, f(X|\mathbf{W},\mathbf{T})P(\mathbf{T}|\mathbf{W})P(\mathbf{W}), \qquad (1)$$

where $(\hat{\mathbf{W}}, \hat{\mathbf{T}})$ expresses the fact that we are looking for a single best template string ($\mathbf{T}$ stands for 'template string') rather than a sum over all template strings explaining a word string. The term $f(X|\mathbf{W},\mathbf{T})$ in equation 1 is the acoustic likelihood of the input given a template string and a word string. It corresponds to the exponent of the negative of the DTW score [4]. The term $P(\mathbf{W})$ is the language model probability and $P(\mathbf{T}|\mathbf{W})$ is the prior probability of a template string given a word string. The condition is binary and resolved through the lexicon, and hence only the prior probability of the template string remains. We can rewrite the template string probability as

$$P(\mathbf{T}) \approx \left\{ \prod_{i=2}^{N_T} P(T_i|T_{i-1}) \right\} P(T_1), \qquad (2)$$

which approximates the probability of a template string by the product of the probabilities of the templates given only their immediate predecessor. When templates are sufficiently long, the influence of the direct predecessor will dominate the influence of the more distant predecessors.

The template transition probabilities can be estimated based on a number of general features, such as acoustic context and information about gender, speaker, environment, etc. In the current system, a simple implementation of equation 2 is used: Different gender of consecutive templates introduces a fixed penalty in the transition probability. This favors paths that are predominantly male or predominantly female. The transition probality is multiplied with a second fixed factor when the original acoustic context of the template differs from the acoustic context in the hypothesis. We use a context length of a single phone. This second penalty can be considered the equivalent of context-dependent HMM models for the example-based approach. Finally, the transition probability is multiplied with a third factor for each concatenation, except when the two templates were neighbors in the original recording. This mechanism favors original sequences of templates longer than a single phone.

2.5. Sentence decoder

A graph decoder combines the HMM phone graph, the template graph, the phonetic lexicon (also a graph, using both prefix and suffix arc sharing), the template concatenation costs and the language model to find the best combined recognition path. We used a simple pseudo-left-to-right strategy with dynamic search space construction and some mild pruning.

We used a linear score combination,

$$S_{combined} = w \cdot D_{DTW} - (1-w) \cdot \log(L_{HMM}), \qquad (3)$$

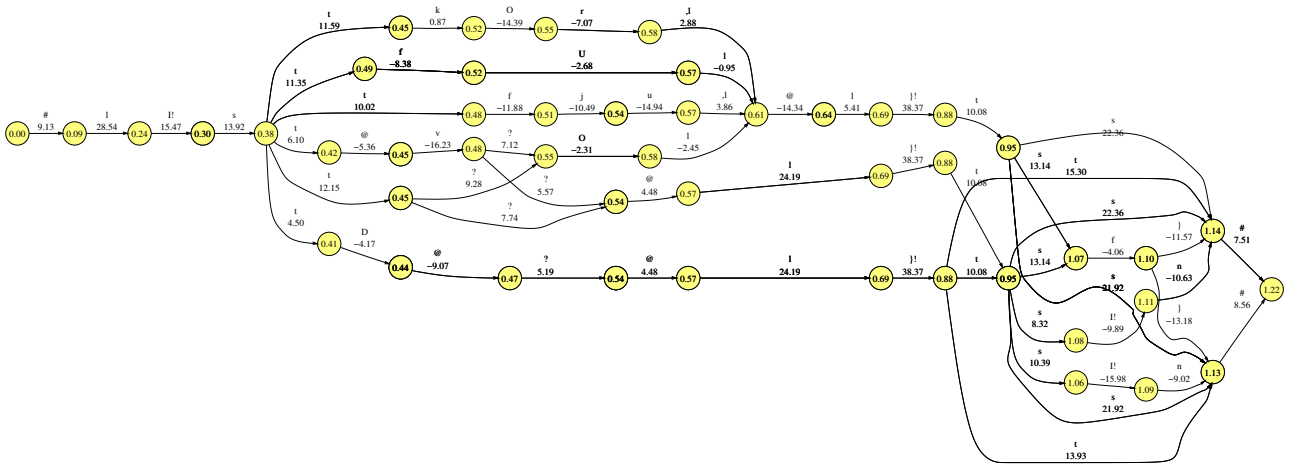

Figure 2: An example of an HMM phone graph.

setup	dev.	oct89	feb91	sep92	mean
HMM only	2.26	2.83	2.13	5.08	3.35
Min. Graph	0.12	0.22	0.08	0.43	0.24
DTW only	3.75	4.02	3.30	6.21	4.51
DTW on HMM	3.05	4.43	3.14	5.35	4.31

Table 1: Baseline word error rates on the different test sets.

i.e. a weighted average of the DTW score D_{DTW} and the negative log HMM likelihood $-\log(L_{HMM})$. Since the template concatenation costs conceptually belong to the DTW side of the recognizer, they were weighted with the same weight as the DTW score.

3. Experimental results

3.1. Task description

All experiments were performed on the Resource Management (RM) benchmark. The RM training database contains about 3.5 hours of noise-free speech. The vocabulary size is 991, and a (non-probabilistic) word-pair grammar is given. We used the CMU v0.4 phonetic lexicon, which has a lot of omissions and mistakes. Improving the lexicon or adding pronunciation rules can dramatically improve recognition accuracy [12], but no effort was made in this respect. Results are presented on four test sets, but the 'feb89' test set was used as a development test set. All average results are calculated over the three other test sets.

3.2. Baseline results

Table 1 shows the relevant baseline results for the combination experiments. 'HMM only' is the baseline HMM result, corresponding to a zero weight w in equation 3. This result compares favorably to long-time reference performance on this benchmark [12].

While 'HMM only' represents the worst possible result we can expect of the combination (in the case the optimal value for w is 0), 'Min. Graph' represents the best possible result we can achieve, as it is the minimal Word Error Rate (WER) of all paths in the graphs. It can be seen that this graph error rate is still far lower than the actual HMM result.

setup	dev.	oct89	feb91	sep92	mean
HMM baseline	2.26	2.83	2.13	5.08	3.35
phone combine	1.87	2.65	1.69	3.52	2.62
no gender	2.11	2.65	2.01	3.66	2.77
no concat cost	2.23	2.94	2.25	4.77	3.32
no DTW score	2.38	2.50	2.13	3.79	2.81

Table 2: Word error rates for HMM baseline, phone-based combination of HMM and DTW and the same without using respectively gender transition cost, all example concatenation costs and DTW acoustic scores.

'DTW only' in table 1 refers to recent results using the bottom-up example-based approach described in [3], using the same features and local scaling as in the other experiments in this paper. 'DTW on HMM' refers to the situation where the weight w in equation 3 is 1. In this case, only DTW scores and concatenation costs are used, but the example graph is based on the HMM graph, which introduces information from the HMM recognizer into the system, especially because the HMM graph is sometimes only one phone 'wide' (see figure2).

3.3. Combination results using phonetic templates

The upper part of table 2 compares the HMM baseline with the result for the combined system using only phone examples (line 'phone combine'). The development test set was used to set the maximum number of templates per HMM arc to 40 and to set the weight w from equation 3 to 0.2. Suitable values for the concatenation costs were also set on the development test set. It can be seen from the table, that the combination approach causes a 22% relative improvement in WER. This leads to a significant improvement at a 99% confidence level, using the non-parametric Bootstrap paired significance test [13].

3.4. Combination results using longer templates

As an extension to the phone-based approach of finding examples for each phone arc in the HMM graph, we also experimented with longer units. From the HMM graph, a graph was computed that contains all biphones in that graph. Based on that biphone HMM

setup	dev.	oct89	feb91	sep92	mean
HMM baseline	2.26	2.83	2.13	5.08	3.35
phone combine	1.87	2.65	1.69	3.52	2.62
mix phone,bi,tri	1.68	2.24	1.65	3.32	2.40

Table 3: WER comparison of phone-based combination and combination based on a mixture of longer templates.

graph, the best matching biphone templates were calculated. The same procedure was used to find the best triphone templates. A mixture graph of phone, biphone and triphone templates was then combined with the original HMM graph. Table 3 shows that a further improvement is possible using this mixture. The presented result was obtained by using a mixture of 20 phone examples, 10 biphone examples and 5 triphone examples if available. The total relative improvement over the HMM baseline is 28%.

3.5. Discussion

The lower part of table 2 sheds some light on the cause of these rather large improvements. 'No gender' is identical to the previous line, except that no gender-based concatenation costs are used. While in the best phone-based combination the chosen template strings are almost entirely gender-consistent (avg. number of gender-consistent phones in sequence is around 30, while the avg. number of phones per recognized sentence is about 38), without gender costs the paths become far less gender-consistent (less than 5 gender-consistent phones in sequence on average). On average adding gender-based concatenation costs results in a 5% relative improvement. 'No concat cost' shows the result of the combination, but *without* using example concatenation costs. Since this result is equal to the HMM baseline, it seems the observed improvement is completely due to the longer-span modeling introduced by the template concatenation costs, rather than a possibly better within-phone acoustic modeling of DTW vs. HMMs. However, setting all DTW scores to the corresponding HMM score and relying on the concatenation costs (line 'no DTW score') also doesn't explain the complete improvement. Therefore, it has to be concluded that the DTW score is useful not only for selecting templates but also for suitably weighting the concatenation costs.

4. Conclusion

In this paper we have introduced a new architecture to efficiently combine HMMs and example-based recognition based on a two-layer approach. Experimental results have shown that the combination outperforms state-of-the-art HMMs. Furthermore, we have pinpointed the reason for the observed improvement: the improvement was mainly due to template concatenation costs, which add long-span information. Thus, we confirmed that HMMs suffer from a lack of long-span modeling, due to the much criticized 'beads-on-a-string' approach. While the experimental results are promising, they are not definitive. On the one hand, the template concatenation model we used was rather primitive, and hence further improvements are certainly possible. On the other hand, the RM benchmark has a good match between training and test sets, which might inflate our results. Experiments on more complex benchmarks that do not exhibit the same limitations as RM (e.g. the Wall Street Journal benchmark) are in progress. Very preliminary results show improvements of 10% relative.

5. Acknowledgments

This research was funded by the Fund for Scientific Research Flanders (FWO-project G.0249.03) and by the IWT in the GBOU programme (project number 020192).

6. References

[1] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. on SAP*, vol. 4, no. 5, pp. 360–378, 1996.

[2] J. A. Bilmes, G. Zweig, T. Richardson, K. Filali, K. Livescu, P. Xu, K. Jackson, Y. Brandman, E. Sandness, E. Holtz, J. Torres, and B. Byrne, "Discriminatively structured graphical models for speech recognition," 2001, Report of the JHU 2001 Summer Workshop.

[3] M. De Wachter, K. Demuynck, D. Van Compernolle, and P. Wambacq, "Data driven example based continuous speech recognition," in *Proc. EUROSPEECH*, Geneva, Switzerland, Sept. 2003, pp. 1133–1136.

[4] M. De Wachter, K. Demuynck, P. Wambacq, and D. Van Compernolle, "A locally weighted distance measure for example based speech recognition," in *Proc. ICASSP*, Montreal, Canada, May 2004, vol. I, pp. 181–184.

[5] M. Matton, M. De Wachter, D. Van Compernolle, and R. Cools, "A discriminative locally weighted distance measure for speaker independent template based speech recognition," in *Proc. ICSLP*, Jeju Island, Korea, Oct. 2004, vol. I, pp. 429–432.

[6] S. Axelrod and B. Maison, "Combination of hidden markov models with dynamic time warping for speech recognition," in *Proc. ICASSP*, Montreal, May 2004, pp. 173–176.

[7] G. Aradilla, J. Vepa, and H Bourlard, "Improving speech recognition using a data-driven approach," in *Proc. EUROSPEECH*, Lisbon, Portugal, Sept. 2005, pp. 3333–3336.

[8] K. Demuynck, T. Laureys, D. Van Compernolle, and H. Van hamme, "Flavor: a flexible architecture for LVCSR," in *Proc. EUROSPEECH*, Geneva, Switzerland, Sept. 2003, pp. 1973–1976.

[9] J. Duchateau, K. Demuynck, and D. Van Compernolle, "Fast and accurate acoustic modelling with semi-continuous HMMs," *Speech Comm.*, vol. 24, no. 1, pp. 5–17, Apr. 1998.

[10] K. Demuynck, J. Duchateau, D. Van Compernolle, and P. Wambacq, "An efficient search space representation for large vocabulary continuous speech recognition," *Speech Comm.*, vol. 30, no. 1, pp. 37–53, Jan. 2000.

[11] K. Demuynck, J. Duchateau, and D. Van Compernolle, "Optimal feature sub-space selection based on discriminant analysis," in *Proc. EUROSPEECH*, Budapest, Hungary, Sept. 1999, vol. III, pp. 1311–1314.

[12] J.-L. Gauvain, L.F. Lamel, G. Adda, and M. Adda-Decker, "Speaker-independent continuous speech dictation," *Speech Communication*, vol. 15, no. 1-2, pp. 21–37, Oct. 1994.

[13] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in asr performance evaluation," in *Proc. ICASSP*, Montreal, Canada, May 2004, pp. 409–412.