

HEAR: An Hybrid Episodic-Abstract speech Recognizer

Sébastien Demange, Dirk Van Compernelle

Katholieke Universiteit Leuven - Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, BELGIUM

{sebastien.demange, dirk.vancompernelle}@esat.kuleuven.be

Abstract

This paper presents a new architecture for automatic continuous speech recognition called HEAR - Hybrid Episodic-Abstract speech Recognizer. HEAR relies on both parametric speech models (HMMs) and episodic memory. We propose an evaluation on the Wall Street Journal corpus, a standard continuous speech recognition task, and compare the results with a state-of-the-art HMM baseline. HEAR is shown to be a viable and a competitive architecture. While the HMMs have been studied and optimized during decades, their performance seems to converge to a limit which is lower than human performance. On the contrary, episodic memory modeling for speech recognition as applied in HEAR offers flexibility to enrich the recognizer with information the HMMs lack. This opportunity as well as future work are exposed in a discussion.

Index Terms: Continuous speech recognition, episodic memory, hybrid architecture

1. Introduction

Almost all current automatic speech recognition (ASR) systems use parametric models and more specifically hidden Markov models (HMMs). These models have been introduced in the mid seventies and they have rapidly become the standard for acoustic modeling in ASR. The HMMs benefit from a well established mathematical formalism, they exhibit a good generalization behavior as well as a compact data representation and efficient automatic training and decoding algorithms are available. Despite major advances over the last quarter of a century, the recognition accuracies seem to converge to a limit which is inferior to human capabilities. The well known HMM weakness is partly responsible for this ceiling of performance. For example, the trajectories and the durations cannot be accurately modeled due to the first order Markov and the state conditional independence assumptions which are unrealistic for speech but needed for computational reasons. Moreover, much useful information, related to fine phonetic detail, is blended with average spectral properties without possible recovery while this is known to play an important role in Human speech recognition [1]. Recently, episodic models have aroused a revival of interest [2, 3, 4]. The underlying recognition paradigm, related to template based or exemplar based speech recognition, considers any linguistic unit encountered in the past as a valuable model for decoding any new speech signal. Instead of building parametric models from past experiences, all past events (referred to episodes, examples or templates) are indexed and stored in memory. For the recognition, the input speech is compared, using a distance, with the templates. The best template sequence is the one which both minimizes the distance to the input speech to be decoded and maximizes its naturalness or consistency w.r.t. predefined criteria. The

transcription is obtained by substituting the templates by the linguistic unit they represent. The major advantage over HMMs is that all the data is preserved intact allowing fine grained acoustic modeling, trajectories and durations retrieval during the recognition. However, such an approach requires much more computational effort.

We propose a hybrid ASR architecture, HEAR, relying on both parametric models, here HMMs, and an episodic memory. HEAR exploits both the HMM capability for producing high quality phone graphs and the opportunity offered by an episodic memory to access fine grained acoustic data. Such a strategy has already been investigated [5, 6, 7]. However, [5, 7] are evaluated on a digit recognition task and the templates correspond to only single digits, and De Wachter in [6] used monophone templates. We propose in this work to use triphones as template units and we evaluate the proposed architecture on a continuous speech recognition task based on a 5k word vocabulary. Such units give opportunity to handle bigger template databases and result in a better acoustic modeling. We show that this architecture is viable and can compete with state-of-the-art HMM based recognizers.

This paper is organized as follow. The proposed architecture is described in section 2. Section 3 presents an evaluation of HEAR as well as a comparison with a state-of-the-art HMM based speech recognizer on the standard Wall Street Journal corpus. Finally, the results are discussed and directions for future works are proposed in section 4.

2. HEAR: overview

HMMs are highly efficient but may fail due to oversimplifications in the model. In contrast, the episodic memory contains all acoustic data about past episodes so that the fine grained information that the HMMs lack can be used to improve the classification. Therefore we use a HMM based recognizer to do the spadework and the episodic memory to refine the classification. In practice HEAR relies on a two-layered decoder architecture.

First, the input signal is processed by a HMM based speech decoder. It results in a dense phone graph containing the more likely recognition hypotheses. Then, for each phone arc within the phone graph we select from the database the N-best matching templates. The timestamps as well as the phone identity of the phone arcs constrain the search. The distance between the input speech segments (aligned with the phone arcs) and each individual template is computed by Dynamic Time Warping (DTW) using the Euclidean distance. A template graph is thus constructed. A template arc is defined by a start time, an end time, a template ID, a phone ID, an

HMM loglikelihood and a DTW score reflecting how close the input signal and the templates are.

Secondly, the template graph is enriched with concatenation cost arcs which are inserted between all consecutive template arcs. They express how natural the concatenation of two templates is. The template graph is then decoded using a finite state transducer combining the lexicon and the language model. This way, the recognition result corresponds to a coherent template sequence which matches the input signal.

2.1. Phone graph generation

This step is crucial for HEAR. Ideally, the phone graph must be as small as possible to reduce the computational time for the template matching. More important, the phone graph must contain a path which corresponds to the true transcription because the template matching will not be able to emit any new hypothesis. The purpose of the template matching is to re-evaluate the HMM recognition hypothesis in the light of new information (episodes). Even if the linguistic unit we use is the phone, the phone graphs are deduced from word graphs rather than being built from a pure phone string recognizer. This way, we prevent any dead-end path (caused by a mismatch with the lexicon) from happening during the decoding and we can introduce stronger top-down knowledge at this early stage to increase the graph quality.

Table 1 shows statistics describing phone graphs obtained with a phone string recognizer and a continuous speech recognizer with different language models. The comparison is made on three graph descriptors: the *graph error* (best WER which can be achieved when considering all paths within the graph), the *density* (average number of context independent phones in parallel per frame) and the *fan out* (average number of arcs leaving a node).

Language model	Phone reco.		Word reco.
	phone 1-gram	phone 4-gram	word 3-gram
Graph. err (%)	1.24	0.69	0.49
Density	5.56	2.78	1.51
Fan-out	4.39	2.96	1.64

Table 1: Comparison between phone graphs produced with different HMM based recognizers on the Dev92 development set.

It is clear that a strong top-down knowledge (here a 3-gram word language model) contributes to significantly improve the graph quality. The graph is more compact and contains most of the time the correct transcription compared with with the two built with the phone string recognizer. Moreover, we have observed that the time alignment of the best path is near optimal with regard to a segmentation using forced alignment of the HMM on the corpus.

2.2. From phone graphs to template graphs

Prior to the decoding, each individual phone hypothesis defined by a phone arc is assigned a DTW score resulting from the comparison of the aligned input signal with the available examples of that phone. It is well known that phonetic contexts strongly influence the realization of a particular phone. De Wachter has used a context mismatch concatenation cost [2] to cope with the

co-articulation effects. It consists in penalizing all templates when their phonetic contexts within the graph differ from their original contexts. In contrary, HEAR uses triphone templates instead of monophone templates as experimented within [4] on a very small vocabulary. Our choice is motivated by many reasons. First the context mismatch concatenation cost looks like a very rigid mechanism because it penalizes identically all context mismatches while different phonetic contexts could result in very similar effects on a particular phone. Second, an episodic memory contains far less examples of a particular triphone than examples of a particular monophone. Thus, it avoids wasting time comparing the input signal with templates whose phonetic contexts make them, a priori, bad examples. Finally, multi-phone modeling is much more flexible so that it is possible to handle longer context dependencies than the strict left and right phones.

In practice, a decision tree is used to define the triphone templates. The splitting of the tree is controlled by phonetic yes or no questions. We query about the left or right context of a triphone, such as “is the left context a plosive?”. A split is done if at least K templates are available for each sub-category and if the acoustic overlap between the sub-categories does not exceed a given threshold. This way, the number of defined triphones is automatically controlled by the size and the richness of the database and in the same time we guarantee that enough examples of each triphones are available. As a consequence, the computational time is kept more or less in bounds. Indeed, as the database grows, the number of triphones increases while the number of examples for each triphone remains stable. So, a much bigger database does not imply much more numerous candidates for the DTW matching. In addition, the partitioning is acoustically consistent since it minimizes the acoustic overlap between the triphones. A particular triphone can then group together different phonetic context configurations if they result in similar realizations of the central phone.

The phone graph obtained at the previous step is context-dependent but consistent with the HMM triphones. So, it is transformed so that each individual phone arc can be mapped to a unique episodic triphone. Once, the N best templates for each phone arc have been selected, the concatenation costs are added. Our goal is to favor long natural template sequences. It means that all template concatenations are assigned a fixed cost except the concatenation of templates which are adjacent within the database.

3. Evaluation

3.1. Experiment setup

The results presented in this paper are obtained using the WSJ (Wall Street Journal) continuous speech recognition corpus. For the evaluation, the 5k close vocabulary development (Dev92) and evaluation (Nov92) sets are used. A trigram word language model is used for both building the phone graphs and decoding the template graphs. Table 2 summarizes the corpora.

The front-end consists of a 36-dimensional feature vector each 10 ms of speech using overlapping frames of 25 ms. First, 24 Mel-scaled filterbank coefficients are extracted. Secondly, a VTLN algorithm is applied and the feature vector is extended by the first and second time derivatives. Finally, MIDA, an improved LDA algorithm, is used to transform the 72 dimensional

	WSJ0+1	Dev92	Nov92
spkr / M / F	284 / 143 / 141	10 / 5 / 5	8 / 4 / 4
length	81 h	46 min	40 min
# sentences	37 516	410	330
# words	643 999	6 780	5 353
# phones	~ 2 800 000	~ 29 000	~ 23 000

Table 2: The WSJ continuous speech recognition task.

space and keep the 36 most informative directions.

Our in house HMM system is made of 15 845 cross-word triphones models and trained on the WSJ0 and WSJ1 training sets (WSJ0+1). The models share 3445 states and each state distribution is made from a pool of 32 725 diagonal covariance Gaussians.

The template database is built from the WSJ0+1 training corpus (284 speakers). It is segmented by forced alignment using our in house HMM system resulting in 2 800 000 phone templates. 4278 episodic triphones are built from a set of 43 monophones as described in section 2.2. Note that the number of episodic triphones is much smaller than the number of HMM triphones. The reason is that even if a particular HMM triphones is rarely observed it is still modeled as the HMMs share their states and Gaussian distributions. On the opposite, the episodic models do not share anything and we do need a minimum of examples of a particular template triphone to reasonably model most of the signal variability. Thus we have fewer but more detailed episodic triphones. An outlier correction procedure called *data sharpening* (DS) [8] is performed on the templates. The acoustic distributions of the templates are sharpened resulting in less overlap between examples of different phones. Finally, the templates are stored in the order they appear in the corpus, so that it is possible to access their preceding and following templates when inserting the concatenation costs.

The context-dependent phone graphs are deduced from word graphs produced by the HMM based speech recognizer as stated in section 2.2. The template graphs are built from the phone graphs by selecting the 50 best matching templates for each phone arc. Note that only a very small fraction of the templates aligned with silence are used during the template selection. Indeed, silence represents a quarter of the database. Moreover, silence does not hold useful acoustic information for speech recognition. Therefore, we prevent the system from wasting time by preselecting silence templates matching the silence arc duration and only 256 of them are compared with the input signal. This way the computational time is reduced by 33% approximately.

3.2. Results

Table 3 summarizes recognition results when the decoder uses only the DTW scores. The purpose of this series of eight experiments is to get a better insight on the advantages and the limits of episodic modeling. Experiments 1 to 4 use monophone templates while experiments 5 to 8 use triphone templates. For each template units four experiments are proposed with combined use of both the *data sharpening* procedure and concatenation costs. For each experiments the word error rate as well as the percentage of natural links is given. We define a natural link as two consecutive templates being part of the recognized template sequence which are adjacent in the template database.

The first remark is that in any configuration the triphone

	WER	% Nat. link
HMM baseline	3.47	-
1) Mono	7.55	0.5
2) Mono + Cost	5.80	23.2
3) Mono + DS	5.49	0.5
4) Mono + DS + Cost	4.68	23.7
5) Tri	5.77	0.6
6) Tri + Cost	4.94	23.8
7) Tri + DS	4.32	3.7
8) Tri + DS + Cost	3.89	28.0

Table 3: Continuous speech recognition results on the Dev92 Wall Street Journal Corpus using only the DTW scores.

templates give better results than the monophone templates. It indicates that a pure acoustic comparison fails to account for the co-articulation effects. Subdividing a broad phone class into many triphones results in better defined set of examples. The overlap between two different phones given their phonetic contexts is then smaller than the overlap between two context-independent phones.

These results also show the lack of generalization inherent in episodic modeling. Indeed, the exemplar based recognition is very sensitive to speech variations. So, HEAR is provided with an efficient generalization/abstraction mechanism. On one hand, *data sharpening* is a bottom-up data normalization. It compensates for unusual speech variability across the episodes by shifting any acoustic feature laying at the tail of an acoustic distribution toward the center of that distribution. On the other hand, the concatenation costs act at the decoding level as top-down constraints. The probability of hearing two identical (or very close) realizations of a particular phone is low, but this probability is even lower when considering the realization of a particular sequence of N phones. Consequently, the DTW distance increases as we consider multi-phone templates. The concatenation costs urge the system to use longer units than single phones by favouring sequences of templates which are adjacent within the database. The experimental results presented in table 3 show how beneficial these abstraction techniques are. Moreover, experiments 4 and 8 show that they're complementary as better recognition results are obtained using both techniques.

The combination of the HMMs and the episodic models is structural as they are embedded together in HEAR. However, this hybrid can also be implemented at the decision level. Indeed, the WER using only the DTW scores in the best configuration (experiment 8) is 12% worse than the HMM baseline but the systems make different errors. For example, the HMM and DTW scores lead respectively to 236 and 264 misrecognized words on the Dev92 set. However, only 176 of these errors are common to both approaches. Hence an ideal combination of both systems could lead to a 25% relative improvement in the WER.

Table 4 presents experimental results using combined scores. We use a simple linear combination of the DTW scores and the HMM loglikelihoods. The concatenations costs are also rescaled but with a different factor. The first and the second rows show that such combination is potentially (because all the system parameters are optimized on the test data) beneficial since the combined scores give the better recognition accuracies. Moreover, the winning template sequences contain much more natural links if we compared the first row with the experi-

ment 8. It indicates that the HMM loglikelihoods contribute further generalization. The third row presents results on the Nov92 set when the system is tuned on the Dev92 set. Here HEAR fails to outperform the HMM baseline. In fact, the Dev92 development set contains records from a female speaker with a strong nasalized voice and the episodic memory fails to provide the system with good examples for this speaker. It underlines the need to normalized the test signal. As discussed in the next section, a modified *data sharpening* approach is conceivable.

Test set	Optimization set	WER		% Nat. link
		HMM	COMB	
Dev92	Dev92	3.47	3.23	37.4
Nov92	Nov92	2.37	2.11	48.1
Nov92	Dev92	2.37	2.48	36.3

Table 4: Continuous speech recognition results on the Dev92 Wall Street Journal Corpus using a combination of the HMM likelihoods and the DTW scores.

4. Discussion and future work

We have shown that HEAR is a viable hybrid ASR architecture with moderate improvement over the HMM baseline. We propose in this section directions for further improvements.

In spite of being very beneficial the score combination is not satisfying. Indeed, we have implicitly supposed, using a fixed linear combination factor, that the recognition errors resulting respectively from the HMM and DTW scores were phone independent. But, how significant the DTW distances are when considering the noise of a fricative or the burst of a plosive? Probably it would be more efficient to use phone-dependent combination weights.

The last experiment suggests that a normalization of the test data could help HEAR to be more robust w.r.t. to the test conditions. *Data sharpening* has been demonstrated efficient for normalizing the templates. This technique shares many similarities with the perceptual magnet effect (PME) [9, 10]. The PME is based on the idea that native language prototypes pull neighboring speech sounds towards them. Perceptual experiments have shown the prototypes to be context-dependent [11]. Moreover, Hintzman, in [12], defined the concept of *echo*. An *echo* is a perceptual representation of an acoustic stimulus built from all stored episodes in memory. The more similar the stimulus and a particular episode are, the more this episode contributes to the *echo*. We think that all these concepts refer to a same abstraction procedure. The episodic models offers the opportunity to implement such mechanism. We intent to investigate this possibility and to provide HEAR with this abstraction procedure prior to the template selection.

5. Conclusion

In this paper we have presented a new ASR architecture, called HEAR, relying on both parametric speech models (HMMs) and episodic memory. This hybrid benefits from the HMM capability for providing high quality phone graphs and the fine grained acoustic modeling offered by the episodic models. HEAR has been evaluated on a standard continuous speech recognition task based on a 5k word vocabulary. The experimental results show a mixed image. In some situations HEAR is able to outperform state-of-the-art HMM systems. On the other hand its

potential is limited by a lack of robustness w.r.t. the test conditions. We believe current results can be further improved by providing HEAR with a perceptual abstraction mechanism.

6. acknowledgment

This work is supported by the Sound-to-Sense project funded by the EU Marie Curie Research Training Network (MC-RTN) and the FWO project G.0260.07 “TELEX” funded by the Flemish Research Foundation.

7. References

- [1] Sarah Hawkins, “Contribution of fine phonetic detail to speech understanding,” in *ICPhS*, Barcelona, Spain, August 2003, pp. 4105–4108.
- [2] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernelle, “Template-based continuous speech recognition,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1377–1390, May 2007.
- [3] Viktoria Maier and Roger Moore, “Temporal episodic memory model: An evolution of minerva2,” in *INTER-SPEECH*, Antwerp, Belgium, August 2007, pp. 866–869.
- [4] V. Ramasubramanian, K. Kulkarni, and B. Kaemmerer, “Acoustic modeling by phoneme templates and modified one-pass dp decoding for continuous speech recognition,” in *ICASSP*, Las Vegas, U.S.A., April 2008, pp. 4105–4108.
- [5] Scott Axelrod and Benot Maisson, “Combination of hidden markov models with dynamic time warping for speech recognition,” in *ICASSP*, Montreal, Canada, May 2004, pp. 173–176.
- [6] M. De Wachter, K. Demuynck, and D. Van Compernelle, “Boosting hmm performance with a memory upgrade,” in *INTER-SPEECH*, Pittsburgh, U.S.A., September 2006, pp. 1730–1733.
- [7] Guillermo Aradilla, Jithendra Jithendra Vepa, and Hervé Bourlard, “Improving speech recognition using a data-driven approach,” in *INTER-SPEECH*, Lisbon, Portugal, September 2005, pp. 3333–3336.
- [8] Mathias De Wachter, Kris Demuynck, and Dirk Van Compernelle, “Outlier correction for local distance measures in example based speech recognition,” in *Proc. ICASSP*, Honolulu, U.S.A., Apr. 2007, vol. IV, pp. 433–436.
- [9] P. Kuhl, “Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not,” *Perception and Psychophysics*, vol. 50, no. 2, pp. 93–107, 1991.
- [10] L. Shi, N. H. Feldman, and T. L. Griffiths, “Performing bayesian inference with exemplar models,” in *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2008.
- [11] S. Hawkins and S. Barret Jones, “The perceptual magnet effect reflects phonetic context,” *The journal of the acoustical society of America*, vol. 115, no. 5, Part 2, pp. 2630, 2004.
- [12] Douglas L. Hintzman, “Schema abstraction in a multiple-trace memory model,” *Psychological review*, vol. 93, no. 4, pp. 421–428, 1986.