Continuous episodic memory based speech recognition using articulatory dynamics

Sébastien Demange, Slim Ouni

Continuous speech recognition using articulatory dynamics

Sébastien Demange

LORIA, UMR 7503
BP329, 54506 Vandoeuvre-lès-Nancy, France
sebastien.demange@loria.fr

Slim Ouni

University Nancy 2 - LORIA, UMR 7503
BP329, 54506 Vandoeuvre-lès-Nancy, France
slim.ouni@loria.fr

Abstract

In this paper we present a speech recognition system which uses articulatory dynamics. We do not extend the acoustic feature with any explicit articulatory measurements but instead the articulatory dynamics of speech are structurally embodied within episodic memories. The proposed recognizer is made of different memories each specialized for a particular articulator. As all the articulators don't contribute equally to the realization of a particular phoneme, the specialized memories don't perform equally regarding each phoneme. We show, through phone string recognition experiments that combining the recognition hypotheses resulting from the different articulatory specialized memories leads to significant recognition improvements.

Index Terms: speech recognition, articulatory dynamic, episodic memory

1. Introduction

Over the last three decades the HMM has been established as the reference models for speech recognition. Its mathematical simplicity, the development of automatic and robust training and decoding algorithms have contributed to its success. However, while important progresses have been done in the past, the recognition accuracies seem to converge asymptotically toward a limit which is situated below the human ability.

This ceiling of performance can be partly explained by some well known HMM limitations. First, though necessary for practical reasons, the HMM probabilistic framework relies on unrealistic assumptions such as the independence of the observations and the first order Markov assumption. Second, it provides insufficient capabilities for modeling trajectories and durations which can be related to the dynamic of speech. Finally, an HMM based speech recognizer usually relies only on two sources of knowledge: the acoustic models of the phonemes and a language model providing the probability to observe a particular word sequence.

Speech perception is much more complex and uses other communication modalities which contribute to disambiguate the acoustic signal. In particular, theories in cognitive science underline the role of the articulation in speech perception [1, 2]. We propose here a new approach relying on episodic memories which make use of articulatory dynamics knowledge. An episodic memory [3] can be viewed as a collection of past events (also called episodes) which have been experienced and stored by a subject. In response to a given stimulus, episodes (similar to the stimulus) are activated in memory and contribute to recognize (categorize) the stimulus. Such a process is expected to take place during speech perception [4]. In addition to its biological basement, an episodic memory is able to describe each episode with respect to different heterogeneous modalities. For example, in this work, we define an episode as a particular realization of a phoneme which has been observed. The realization is described by both acoustic and articulatory observation sequences.

Recently, databases of synchronized acoustic and articulatory data streams using electromagnetic articulography have been possible. However, due to practical difficulties during the data acquisition, most of them is limited to few tens of minutes. Then, only few episodes of each phoneme are available which is quite insufficient for covering the important speech variability. The originality of the proposed memory lies in its ability to combine different episodes during the recognition process based on their articulatory dynamics. The memory is extended with inter-episodes transitions based on articulatory continuity constraints and which respect the original temporal order of the observations. Any path across different episodes of a same phoneme result to a particular acoustic realization which would have been produced by the resulting articulatory observation sequence.

In the next section we present how the memory is built. Section 3 explains how to recover the phone string from an acoustic speech signal using the acoustic-articulatory memory. The corpora, the feature extraction as well as the experiments set-up are presented in section 4 and the recognition results are exposed along the section 5.

2. Acoustic-articulatory episodic memory

The main advantage of an episodic memory is that it keeps trace of the order of the observations and thus preserves the acoustic and articulatory dynamics of each episode. In order to preserve this property, the inter-episodes transitions have to be defined carefully. Indeed, they're expected to provide the memory with generalization capabilities but must not allow the memory to produce unrealistic pattern from a dynamical point of view. The inter-episode transitions are defined accordingly to the concept of articulatory target interval (ATI).

2.1. Articulatory target interval

Let X be a particular articulatory realization of a given phoneme expressed as a sequence of K observations: $X = (x_1, x_2, \ldots, x_K)$. We define each observation x_{i+1} as the natural articulatory target of x_i as it has been observed following x_i. In fact, x_{i+1} is a particular articulatory configuration but we can suppose it could have been slightly different. Indeed, starting from x_i at time t, the articulators could have reached a different target at time $t+1$ close to x_{i+1} without affecting the rest of the realization. Then, for each x_i we define an articulatory target interval ATI_{x_i} as the interval $[x_{i+1} - \delta, x_{i+1} + \delta]$, where δ is a positive value.

2.2. Modeling the articulatory dynamics

2.2.1. Inter-episode transitions

Let $Y = (y_1, y_2, \ldots, y_N)$ be a second articulatory realization of the same phoneme as X. We define $\phi = (\Phi_1, \ldots, \Phi_M)$ as the alignment path corresponding to the shortest distance $D(X, Y)$ between X and Y obtained by the well-known dynamic time warping algorithm (DTW). Each Φ_i is a pair of indexes of the elements of X and Y, which are aligned together: $\Phi_i = (\Phi_{x,i}, \Phi_{y,i})$. For example, $\Phi_3 = (4, 5)$ indicates that the third element of the path is an alignment between x_4 the fourth articulatory configuration of X and y_5 the fifth articulatory configuration of Y. For our problem, we extended the DTW algorithm with the Itakura constraints [5] to impose temporal constraints on the alignment paths ensuring that aligned articulatory configurations occur at similar time in their respective episode. Once the DTW distance between X and Y is computed, a transition in the memory from any x_i to a y_j is created if y_j matches the two following conditions:

$$\Phi_{y,i+1} = j \quad (1)$$

$$y_j \in ATI_{x_i} \quad (2)$$

Equation 1 requires y_j to be aligned with x_{i+1} when mapping Y onto X . In other words, it has to be aligned with the following observation of x_i. This condition ensures that the transitions are consistent with the temporality of the episodes. Equation 2 states that y_j has to belong to the $ATIx_i$. It locally ensures the physical articulatory validity and naturalness of the transition since y_j is close to x_{i+1}, which is the natural articulatory target of x_i. Note that the articulatory trajectories of two episodes of the same phone can be significantly different due to the coarticulation effects as their phonetic contexts can differ. Combining two episodes, which match only on a very small segment but which drastically differ outside, could result in unrealistic trajectory. To avoid this undesired effect, transitions from X to Y are created only if Y is similar enough to X:

$$D(X, Y) \leq \Delta \quad (3)$$

where Δ is a positive value. One can remark that the memory is still conservative as all the original episodes it is made of are preserved since:

$$D(X, X) = 0 \leq \Delta \quad (4)$$

$$\Phi_{x,i+1} = i + 1 \quad (5)$$

$$x_{i+1} \in ATI_{x_i} \quad (6)$$

2.2.2. Between-episode transitions

The transitions at episode boundaries are only subject to the articulatory continuity requirement expressed by the equation 2. Let Z be the episode observed following X. Then, a transition from x_K (the last observation of X) to the first observation w_1 of any realization W of any phoneme is created if $w_1 \in ATI_{x_K} = [z_1 - \delta, z_1 + \delta]$. If the episode X is the last of a record the natural articulatory target of x_K is unknown and equation 2 cannot be verified, thus no transition to any other episode is possible from x_K.

3. Phone string recognition

In practice the memory is modeled as an oriented graph. The nodes are synchronized acoustic and articulatory observations and the edges are the allowed transitions reflecting natural articulatory dynamics. The transitions are created according to the procedure described above in the articulatory space using Euclidean distance. Each path within the graph corresponds to a physically possible articulatory trajectory and its corresponding acoustic counterpart. In addition, the nodes are supplemented with linguistic and temporal attributes which are the phoneme label of the acoustic-articulatory observations and their relative index within their original episode. We define the relative index $RI(x_i)$ of any observation x_i being part of an episode X made of K observations as i/K.

Recognizing speech consists in finding the path within the memory which acoustically best matches the speech signal X to be recognized. All paths can start only at nodes representing the first observation of the original episodes and can end only at nodes representing the last observation of the original episodes. During the recognition a breadth first search is performed applying the Viterbi algorithm. At each step, only the K best paths are propagated through all defined transitions. The score of each path is the sum of the local acoustic distances between the visited nodes and the observations of X. The local acoustic distances are computed over a window using the Euclidean distance.

The recognized phone string is deduced from the winning path. The phone segmentation is deduced from the evolution of the RI of each node along the path. Any part of the path across different episodes of the same phoneme might exhibit a continuous increase of the RIs as the inter-episodes transitions respect the temporality of the original episodes. A significant decrease indicates that the path start a new phoneme as the transitions at the episode boundaries are defined between the last ($RI = 1$) and the first (small RI) observations of the original episodes. All the observations within a particular segment comes from realizations of the same phoneme, that is they do not necessarily come from a unique episode, but may come from multiple episodes of a same phoneme between which inter-episode transitions have be defined.

Table 1: Synthetic description of the corpora.

Corpora	Sets	Durations	Sentences	Phones
fsew	train	16 min 35 sec	368	11179
	dev	1 min 57 sec	46	1324
	test	2 min 5 sec	46	1457
msak	train	13 min 59 sec	368	11179
	dev	1 min 41 sec	46	1324
	test	1 min 45 sec	46	1457
mdem	train	8 min 24 sec	319	6355
	dev	1 min 2 sec	40	817
	test	1 min 3 sec	40	814

4. Experiment set-up and corpora

4.1. Corpora

All the experiments presented in this work have been carried out on two corpora of synchronized acoustic speech signal and articulatory trajectories. Table 1 synthesizes the corpora.

The first corpus is MOCHA [6]. Two speakers, a female (*fsew*) and a male (*msak*) British English speakers, were recorded while reading 460 short phonetically balanced British-TIMIT sentences. We use in this work the acoustic and EMA data streams. The acoustic is provided as waveforms sampled at

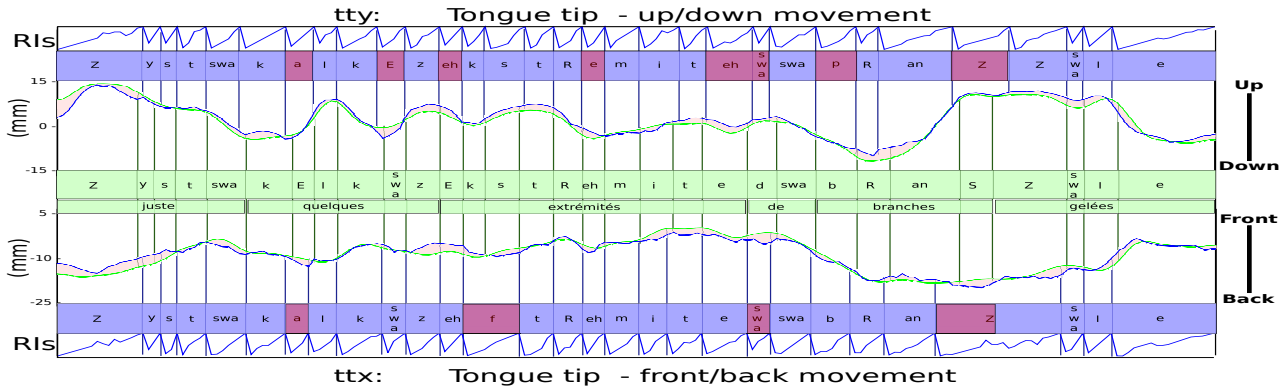

Figure 1: Tongue tip specialized memories recognition hypotheses resulting from the decoding of the French sentence “juste quelques extrémités de branches gelées”.

16 kHz and the EMA data consist in 2D data with coordinates expressed in the mid-sagital plane. Nine sensors are used, located at the bridge of the nose (bn), upper incisors (ui), lower incisors (li), upper lip (ul), lower lip (ll), tongue tip (tt), tongue body (tb), tongue dorsum (td) and velum. The two first are used to normalize the trajectories of the last seven with regard to the head movements. The y axis (upright) passes through the upper incisors (origin of the coordinate system) and bridge nose sensors.

The second one is a corpus we have recorded with an articulograph (AG500, Carstens Medizinelektronik). For now only one speaker, a male French speaker (*mdem*) has been recorded reading 400 phonetically balanced sentences. The data streams are synchronized acoustic waveforms sampled at 16 kHz and 3D EMA data. We used 6 sensors fixed in the mid-sagital plane on the lower lip (ll), upper lip (ul), tongue tip (tt), tongue body (tb), tongue dorsum (td) and tongue back dorsum (tbd). Three additional sensors have been used for normalizing the articulatory trajectories with regard to the head movements, one at the bridge of the nose and two located behind each ear. Prior to the features extraction the EMA data axis have been shift and rotated so that each articulatory sample is expressed in the 2D mid-sagital plane. The phonemic segmentation has been obtained by forced aligning French acoustic models onto the acoustic stream.

4.2. Features extraction

The silences occurring at the beginning and the end of the records are first discarded as the articulators may move unpredictably. A Linear Predictive Analysis [7] is performed on the speech signal using the HTK toolkit [8]. 12 cepliftered MF-PLPs [9] plus the logarithmic energy of the signal comprise the acoustic feature vector extracted from every 25 ms speech frame shifted by 10 ms.
The articulatory data are first down sampled to 100 Hz to match the acoustic frameshift. Then, all trajectories are low-pass filtered in order to remove the recording noise using a cut-off frequency of 20 Hz.

4.3. Experiments set-up

The memory parameters have been optimized on the development sets. For each corpus, a memory is built for each tracked articulator and along both the horizontal and vertical axis. That is, each memory can be considered as an articulatory specialized speech recognizer. Each memory provides a phone string recognition hypothesis. The final recognition result is obtained combining these hypotheses using a majority vote and with the constraint that all phone segments within the resulting phone transcription is at least 30 ms long.

5. Results

Figure 1 shows an example of the memory outputs specialized on the tongue tip dynamics. The acoustic signal to be recognized is the French sentence “juste quelques extrémités de branches gelées” which can be translated by “only few frozen extremities of tree branches”. The reference word and phone segmentations are provided at the middle of the figure. The upper and the lower parts, plot the relative index of the nodes along the winning decoding paths, the respective phone string segmentations (with misrecognized phoneme in red) and the articulatory trajectories (in blue) obtained from the articulatory observations of the nodes along the winning paths. For comparison the reference articulatory trajectories are provided with the green curves.

The articulatory specialized phone transcriptions are impressive knowing that no linguistic information (such as phone language model, or word dictionary) is provided and knowing that the memories are based only on few minutes of speech. Only, the acoustic distances computed on the static acoustic parameters and the articulatory dynamics contribute to the transcriptions. As expected, each memory performs differently. Let’s consider the /k/ and /s/ which occur at the beginning of the word “extrémités”. The places of articulatory constrictions are between the tongue body and the palate for the /k/ and the tongue tip and the palate just behind the upper incisors for the /s/. The realization of these two phones implies a raise of the tongue tip while the tongue tip doesn’t move significantly along the front/back axis. The acoustic realization consists in a occlusion followed by a released of the air stream for the /k/ and fricative noise for the /s/. Note now, that the tt_y specialized memory succeeds in recognizing this phoneme sequence while the tt_x fails. The tt_x memory has recognized a /f/ which is fricative noise from an acoustic point of view. Regarding the tongue tip, the articulatory realization of a /f/ is close to the realization of a /s/ but the tongue tip doesn’t raise in order to facilitate the air flow through the mouth as the constriction is between the upper incisors and the lower lip. Therefore, the movement of the tongue tip along the up/down axis is critical

Table 2: Phone recognition error rates (PER) in %.

	mdem	*fsew*	*msak*
lix	-	41.3	49.0
liy	-	45.7	44.9
llx	29.6	42.5	45.9
lly	27.3	44.1	47.9
ulx	27.1	43.4	53.2
uly	26.5	46.4	44.4
ttx	27.0	41.2	41.9
tty	26.8	45.7	44.5
tbx	31.2	43.0	44.8
tby	25.5	45.9	52.6
tdx	25.0	42.0	43.1
tdy	29.6	45.8	45.2
tbdx	27.3	-	-
tbdy	29.9	-	-
vlx	-	43.3	46.5
vly	-	47.8	53.2
Average	27.7	44.2	46.9
Combination	21.8	36.3	39.7
HMM	22.8	36.0*	38.0*

for distinguishing between the /s/ and the /f/. This is a typical example of how the phone recognition results differ across the articulatory specialized memories. Then combining the different recognition hypothesis should result in an improved phone error rate.

Table 2 summarizes the recognition results obtained from each specialized memory as well as the PER averaged over all the memories. The last two rows give PER after the majority vote combination as well as pure acoustic HMM references. For *mdem* we have trained left-to-right, three states monophone HMMs, each state modeled by a mixture of 8 Gaussian. For MOCHA we didn't train HMMs, instead we use the HMM baseline from [10] which match our frontend. Their models consist in left-to-right three states triphone HMMs, each state modeled by a mixture of 2 to 7 Gaussian. In addition they used a bi-gram phone language model. Contrary to the memories, the HMM acoustic feature vectors are supplemented with the first and second time derivatives.

The results show that all the specialized memories give similar PER. The combination is very useful, resulting in a PER reduction ranging between 15% and 20% according to the corpora. Moreover, the PERs resulting from the combination are similar to the HMM baselines.

We compare our PER results with those published by Frankrel and King [11] as our experiment share the same MOCHA corpus and our decoding experiements both take place in the context of phone string recognition. They obtained PERs of 33% in a classification experiment and 44% in recognition experiment while our average PER on the same dataset is 38%. The authors supplemented the acoustic feature vectors with articulatory feature estimated using a recurrent neural network. Contrary to Frankrel and King we do not make explicit use of articulatory measurements, but rather the memory structurally embody the articulatory dynamics. Such a strategy prevent from propagating acoustic-to-articulatory inversion errors to the recognizer.

6. Conclusions

We have proposed a new episodic memory based approach for speech recognition. The memory is able to combine different episodes it is made of based on their articulatory dynamics. We have shown that the articulatory dynamics strongly influence the recognition results as all articulatory specialized memories perform differently. Combining the articulatory specialized recognition hypotheses results in significant improvements.

Though impressive, regarding the limited size of the training material and the absence of linguistic information, the recognition results are still preliminary and could be significantly improved. We use the simple Euclidean distance for computing the acoustic distances during the decoding and this distance is known not to be robust for speech. The use of a local kernel based distance [12] would improve the recognition. Moreover our acoustic frontend consists in only static acoustic features while the first and second derivatives would provide information of the acoustic dynamics. Finally, our majority vote combination is very simple. It seems more efficient to develop a combination procedure based on the notion of critical articulator. The weights of each articulatory specialized memory would be dynamically computed regarding the hypothesized phonemes at time. For a given hypothesized phoneme P memories which are specialized on the critical articulators for the phoneme P should contribute more to the combination.

7. References

[1] L. Fadiga, L. Craighero, G. Buccino, and G. Rizzolati, "Speech listening specifically modulates the excitability of tongue muscles: a tms study," *European Journal of Neuroscience*, vol. 15, no. 2, pp. 399–402, January 2002.

[2] K.E. Watkins, A.P. Strafella, and T. Paus, "Seeing and hearing speech excites the motor system involved in speech production," *Neuropsychologia*, vol. 41, no. 8, pp. 989–994, 2003.

[3] E. Tulving, "Episodic and semantic memory," *Organization of Memory*, pp. 381–402, 1972.

[4] S.D. Goldinger, "Echo of echoes? an episodic theory of lexical access," *Psychological Review*, vol. 105, no. 2, pp. 251–279, 1998.

[5] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 67–72, February 1975.

[6] A.A. Wrench and W.J. Hardcastle, "A multichannel articulatory database and its applicationfor automatic speech recognition," in *5 th Seminar of Speech Production*, Kloster Seeon, Bavaria, May 2000, pp. 205–308.

[7] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, April 1990.

[8] S. Young, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK version 3.3)*, Cambridge University Engineering Department, 2005.

[9] P. Woodland, M. Gales, Pye D., and S. Young, "Broadcast news transcription using htk," in *ICASSP*, Washington DC, USA, 1997, IEEE.

[10] A.A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," in *ICSLP*, 2000, pp. 145–148.

[11] J. Frankel and S. King, "Asr - articulatory speech recognition," in *Eurospeech*, 2001, pp. 145–148.

[12] M. De Wachter, *Example Based Continuous Speech Recognition*, Ph.D. thesis, Katholieke Universiteit Leuven, ESAT, May 2007.