Continuous episodic memory based speech recognition using articulatory dynamics

Sébastien Demange\textsuperscript{1}, Slim Ouni\textsuperscript{2}

\textsuperscript{1}INRIA, \textsuperscript{2}University of Nancy 2 - LORIA, UMR 7503
BP239, 54506 Vandoeuvre-lès-Nancy, France
sebastien.demange@loria.fr, slim.ouni@loria.fr

Abstract
In this paper we present a speech recognition system based on 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 do not contribute equally to the realization of a particular phoneme, the specialized memories do not 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 dynamics, 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 achieved, the recognition accuracies seem to converge asymptotically toward a limit which is situated below the human performances.

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 dynamics of speech. Finally, an HMM based speech recognizer usually relies only on two sources of knowledge: acoustic models of 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, research in neuroscience underlines 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 human. 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 [5]. For example, in our 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 available. However, due to practical difficulties during data acquisition, most of them are limited to few tens of minutes. Then, only few episodes of each phoneme are available which is quite insufficient to cover the important speech variability. The originality of the proposed memory is 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 results 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 setup are presented in section 4 and the recognition results are exposed along section 5.

2. Acoustic-articulatory episodic memory
The main advantage of an episodic memory is that it keeps track 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 are expected to provide the memory with generalization capabilities but must not allow the memory to produce unrealistic pattern from a dynamic 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: \( X = (x_1, x_2, \ldots, x_K) \). The observations are articulatory configurations which have been measured at a fixed sampling rate. 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 that it could be slightly different. Indeed, starting from \( x_i \) at time \( i \), the articulators could have reached a different target at time \( i + 1 \).
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 \( \delta \) is a given positive value.

2.2. Modeling 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 \( \Phi_i \) 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 [6] to impose temporal constraints on the alignment paths ensuring that aligned articulatory configurations occur at equivalent instants in their respective episode. Once the DTW distance between \( X \) and \( Y \) is computed, a transition in the memory from any \( x_i \) to any \( y_j \) is created if \( y_j \) matches the two following conditions:

\[
\Phi_{y,i+1} = j \\
y_j \in ATI_{x_i}
\]

Condition 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. Condition 2 states that \( y_j \) has to belong to the \( ATI_{x_i} \). 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 co-articulation 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 \( X \):

\[
D(X, Y) \leq \Delta
\]

where \( \Delta \) 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:

\[
D(X, X) = 0 \leq \Delta
\]

\[
\Phi_{x,i+1} = i + 1
\]

\[
x_{i+1} \in ATI_{x_i}
\]

2.2.2. Between-episode transitions

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

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 its position within this episode: \( i/K \).

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 natural episodes and can end only at nodes representing the last observation of the natural 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 phone segmentation is inferred 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 \( RI \)'s 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, i.e. 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 been defined.

4. Corpora and experiment setup

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 split within train, development and test sets.

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

The first corpus is MOCHA [7]. Two speakers, a female (fsew) and a male (msak) British English speakers, were recorded while reading 460 short phonetically balanced British-
Figure 1: Recognition hypotheses resulting from the decoding of the French sentence “juste quelques extrémités de branches gelées” by memories specialized on the tongue tip along the up/down and front/back axes. The reference word and phone segmentations are provided in the middle (in green). The upper and lower parts show the plots of the RI, the recognized phone strings (misrecognized phonemes in red) as well as the plots of the reference articulatory trajectories (green) and articulatory trajectories recovered from the winning decoding path (blue).

TIMIT sentences. We use in this work the acoustic and EMA data streams. The acoustic is provided as waveforms sampled at 16 kHz and the EMA data consist in 2D data with coordinates expressed in the mid-sagittal 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 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). 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-sagittal 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 to remove head movements, one at the bridge of the nose and two located behind each ear. Prior to feature extraction the EMA data axis have been shift and rotated so that each articulatory sample is expressed in the 2D mid-sagittal plane. The phonetic segmentation has been obtained by forced aligning French acoustic models onto the acoustic stream.

4.2. Feature extraction

The silences occurring at the beginning and the end of the records are first discarded as articulators may move unpredictably. A Linear Predictive Analysis is performed on the speech signal using the HTK toolkit.12 cepliftered MF-PLPs 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 setup

The memory parameters are optimized on the development sets minimizing the error of the estimated articulatory trajectories. The $\delta$ and $\Delta$ thresholds were optimized jointly through a grid search. For each corpus, a memory is built for each tracked articulator and along both the horizontal and vertical axes. 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 by 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 in the middle of the figure. The upper and 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. It is interesting to note that even when a memory misrecognizes a particular phoneme the articulatory trajectory, recovered from the winning decoding path, is still very close to the reference articulator trajectory. This indicates that the articulator for which the memory is specialized is not appropriated to distinguish between the recognized and the reference phonemes.

Table 2 summarizes the recognition results obtained by each of the specialized memory as well as the phone error rate (PER) averaged over all the memories. The last two rows give PER after the majority vote combination as well as PER of our acoustic HMM baselines. For \( m_{\text{test}} \) the models are left-to-right, three states monophone HMMs, each state modeled by a mixture of 8 Gaussian. For MOCHA the models consist in left-to-right three states triphone HMMs, each state modeled by a mixture of 2 to 7 Gaussian. In addition a bi-gram phone language model is used during the decoding. Unlike the memories, the HMM acoustic feature vectors are supplemented with the first and second time derivatives. The HMM are trained on several hours of generic speech and then adapted to the speaker.

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, although the memories are based only on few minutes of speech, the PERs resulting from the combination are similar to the HMM baselines. We compare our PER results with those of Frankel and King [8] as our experiment share the same MOCHA corpus and our decoding experiments 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 Frankel and King we do not make explicit use of articulatory measurements, but rather the memory structurally embodies the articulatory dynamics. Our strategy prevents from propagating articulatory estimation errors to the recognizer.

6. Conclusions

We have proposed a new episodic memory based approach to speech recognition integrating articulatory dynamics. The memory is able to combine natural episodes based on their articulatory dynamics. We have shown that the articulatory dynamics strongly influence the recognition results as the 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 during decoding, the recognition results are still preliminary and can be significantly improved. We use a simple Euclidean distance to compute the acoustic distances during the decoding and this distance is known not to be robust for speech signals. The use of a local kernel based distance [9] would improve the recognition. Moreover, our acoustic frontend consists in only static acoustic features while the first and second derivatives would provide information on 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 each time. Memories which are specialized on the critical articulators of a given phoneme should contribute more to the combination when this phoneme is hypothesized.

7. References