HANDLING DISYNCHRONIZATION PHENOMENA WITH HMM IN CONNECTED SPEECH

Pierre Jourlin
Laboratoire d'Informatique, C.E.R.I
339, Chemin des Meinajariès, BP 1228, 84911 Avignon Cedex 9, France
Tel: +33 90 84 35 35; fax: +33 90 84 35 01
e-mail: jourlin@univ-avignon.fr

ABSTRACT

Anticipation and retention phenomena between the different phonatory organs have been widely studied in the speech perception and production domain. However, few automatic speech recognition systems are able to handle them. In this paper, we define a product of valuated transitions automata handling these difficulties. Then, we use such automata in a recognition system based on HMM. This method is evaluated in two different contexts : bimodal and unimodal speech recognition. The results show an improvement for the the product model against a synchronous one of 1.9% in the bimodal field and of 1.2% in the unimodal one.

1 INTRODUCTION

The multimodal aspect of the speech perception process has been widely studied. During the last decade, some automatic bimodal speech recognition systems have been designed [1, 2]. Anticipation and retention phenomena between the different phonatory organs have been widely studied in the speech perception and production domain [3], but few recognition systems are assumed to take into account the anticipation and retention phenomena between acoustics and lip movement [4, 5]. In order to cope with such problems, the aim of this paper is to define a product applied to two valuated-transition automata related to two different sources of information. These definitions are applied to continuous HMM, in order to merge an acoustic model and a labial model. Such a model, which we call acoustico-labial, is well-suited to handle the anticipation and retention phenomena. Furthermore, if we can split the acoustic space into several unsynchronized subspaces for each recognition unit, then we can use this method to handle disynchronization in the acoustic source. So far, we have only managed to split the acoustic space into two subspaces in an arbitrary way and without taking into account the different recognition units. Using these two subspaces, we have built a set of purely acoustic product-models. We compare product models and synchronous models in the bimodal and in the unimodal speech fields. This work is supported by the AMIBE project [6].

2 PRODUCT OF TWO VTA

2.1 Definition of a VTA

A valuated transition automaton (VTA) is entirely defined by $(S, \mathcal{O}, (I, \cdot), p, d)$, where :

- $\mathcal{O}$ is any finite set (called observation set),
- S is any finite set (called state set),
- $(I, \cdot)$ is an commutative semi-group (called probability set),
- L is the set of functions from $\mathcal{O}$ into I (called distribution set),
- p is a function from S^2 into I (called transition function),
- d is a function from S into L (called distribution function).

2.2 Definition of the product of two VTA

Let A and A' be two VTA, respectively defined by $(S, \mathcal{O}, (I, \cdot), p, d)$ and $(S', \mathcal{O}', (I, \cdot), p', d')$. We call product of A and A' the operation " $*$ " such that $A'' = A * A'$ defined by $(S'', \mathcal{O}'', (I, \cdot), p'', d'')$, where :
$S'' = S \times S'$, $\mathcal{O}'' = \mathcal{O} \times \mathcal{O}'$, $p'' : (S \times S')^2 \rightarrow I$, $d'' : S \times S' \rightarrow L''$.
L'' is the set of functions from $\mathcal{O} \times \mathcal{O}'$ into I.
and :
$p''(((s_w, s'_x), (s_y, s'_z))) = p((s_w, s_y)) \cdot p'((s'_x, s'_z))$,
$\forall (s_w, s_y) \in S^2$, $\forall (s'_x, s'_z) \in S'^2$
$d''((s, s')) = l'' \in L''$ such that $l''((o, o')) = l(o) \cdot l'(o')$,
where $l = d(s)$ and $l' = d'(s')$ $\forall (s, s') \in S \times S'$, $\forall (o, o') \in \mathcal{O} \times \mathcal{O}'$

A'' is a VTA.

3 EVALUATION OF SYMBOL SEQUENCES PRODUCED BY FOLLOWING A PATH IN A VTA

Let (C) be a set of sequences of elements of $S \times \mathcal{O}$. We call evaluation of a sequence of elements of $S \times \mathcal{O}$ in a VTA $A : (S, \mathcal{O}, (I, \cdot), p, d)$ the function f such that :

$$f : A \times (C) \rightarrow I$$

$$f(S,\mathcal{O},(I,\cdot),p,d,(c)) = d_{s_n}(o_n) \cdot \prod_{i=1}^{n-1} (d_{s_i}(o_i) \cdot p(s_i, s_{i+1}))$$

$n = |(c)|$, $(s_i, o_i) \in (c)$, $s_i \in S$, $o_i \in \mathcal{O}$, $\forall i \in [1,n]$.

3.1 Properties

Let $A(S,\mathcal{O},(I,\cdot),p,d)$, $A'(S',\mathcal{O}',(I,\cdot),p',d')$ and $A''(S'',\mathcal{O}'',(I,\cdot),p'',d'')$ be three VTA such that $A'' = A * A'$. Considering now the set (C'') of couples of paths $((c),(c'))$ in $(C) \times (C')$ of the same length.

Property 1

$$\forall ((c),(c')) \in (C'') \; f(A,(c)) \cdot f(A',(c')) = f(A'',(c''))$$

$\forall i \in [1,n] \; : c''_i = ((s_i,s'_i),(o_i,o'_i)), (s_i,o_i) \in (c)$ *and* $(s'_i,o'_i) \in (c')$

Demonstration 1

$$\begin{aligned}
& f(A,(c)) \cdot f(A',(c')) \\
= \; & d_{s_n}(o_n) \cdot \prod_{i=1}^{n-1} (d_{s_i}(o_i) \cdot p(s_i,s_{i+1})) \\
& \cdot d'_{s'_n}(o'_n) \cdot \prod_{i=1}^{n-1} (d'_{s'_i}(o'_i) \cdot p'(s'_i,s'_{i+1})) \\
& (s_i,o_i) \in (c), (s'_i,o'_i) \in (c'), \forall i \in [1,n] \\
& \textit{As } \cdot \textit{ is associative and commutative :} \\
= \; & d_{s_n}(o_n) \cdot d'_{s'_n}(o'_n) \cdot \prod_{i=1}^{n-1} d_{s_i}(o_i) \\
& \cdot d'_{s'_i}(o'_i) \cdot p(s_i,s_{i+1}) \cdot p'(s'_i,s'_{i+1}) \\
& \textit{By definition of } d'' \textit{ and } p'' : \\
= \; & d''_{(s_n,s'_n)}((o_n,o'_n)) \cdot \prod_{i=1}^{n-1} (d''_{(s_i,s'_i)}((o_i,o'_i)) \\
& \cdot p''((s_i,s'_i),(s_{i+1},s'_{i+1}))) \\
& \textit{By definition of } (c'') : \\
= \; & f(A'',(c''))
\end{aligned}$$

4 EVALUATION OF A SEQUENCE OF SYMBOLS PRODUCED BY A VTA

4.1 Definition

Let (O) be the set of sequences of elements of $\mathcal{O}$ and (C) the set of sequences of elements of $S \times \mathcal{O}$. Let $(C)_{(o)}$ be the subset of sequences of (C) such that if (s_j,o_j) is the j^{th} element of $(C)_{(o)}$ then o_j is the j^{th} element of (o).

We call evaluation of a sequence of elements of (O) for a given VTA, the function g such that :

$$g : (S,\mathcal{O},(I,\cdot),p,d) \times (O) \rightarrow I$$

$$g(S,\mathcal{O},(I,\cdot),p,d,(o)) = \sum_{(c)\in(C)_{(o)}} f(S,\mathcal{O},(I,\cdot),p,d,(c))$$

4.2 Properties

Property 2 *Let A, A' and A'' be three VTA such that $A'' = A * A'$:*

$$g(A,(o)) \cdot g(A',(o')) \; = \; g(A'',(o''))$$

$$\forall ((o),(o')) \in (O) \times (O') \textit{ such that } |(o)| = |(o')|$$

Demonstration 2

$$\begin{aligned}
& g(A,(o)) \cdot g(A',(o')) \\
= \; & \Big(\sum_{(c)\in(C)_{(o)}} f(A,(c))\Big) \cdot \Big(\sum_{(c')\in(C')_{(o')}} f(A',(c'))\Big) \\
= \; & \sum_{(c)\in(C)_{(o)},(c')\in(C')_{(o')}} f(A,(c))) \cdot f(A',(c')) \\
& f \textit{ verifying property 1 :} \\
= \; & \sum_{(c'')\in(C'')_{(o'')}} f(A'',(c'))) \\
= \; & g(A'',(o''))
\end{aligned}$$

5 APPLICATION TO CONTINUOUS MARKOV MODELS

A Markov model being a finite state automaton, we can substitute :

- $\mathcal{O}$ by $\mathcal{R}^n$
- S by the set of states of the model.
- $(I,\cdot)$ by the commutative semi-group $([0,1],\cdot)$ subset of $\mathcal{R}$.
- L by the set of probability distributions.
- p by the function which associates a probability to each transition.
- d by the function which associates a probability distribution to each emitting state.
- f by the function which calculates the probability of emitting a given sequence of observation following a given path in a given model.
- g by the function which estimates the *a posteriori* probability of emitting a given sequence of observations with a given model.

A continuous hidden Markov model (CHMM) is entirely defined by (S,p,d).

Let :

- $A(S,p,d)$, $A'(S',p',d')$, $A''(S'',p'',d'')$, be three CHMM such that $A'' = A * A'$ (we can show that $\forall s''_1, s''_2 \in S'', p''(s''_1,s''_2) > 0$ and $\forall s''_1 \in S'', \sum_{s''_2 \in S''} p''(s''_1,s''_2) = 1$)

- (o) be a sequence of vectors of $\mathcal{O}$ and (o') a sequence of vectors of $\mathcal{O}'$ such that $|(o)| = |(o')|$.
- (s) be a path of state of A and (s') a path of state of A'.
- (c), (c'), (c'') such that $c_i = (s_i, o_i)$, $c'_i = (s'_i, o'_i)$ and $c''_i = ((s_i, s'_i), (o_i, o'_i))$ where $c_i \in (c)$, $c'_i \in (c')$ and $c'_i \in (c')$ for all i in $[1, |(o)|]$.

By definition of the product of two VTA, we have the following properties :

$$\begin{aligned} f(A,(c)) \cdot f(A',(c')) &= f(A'',(c'')) \\ g(A,(o)) \cdot g(A',(o')) &= g(A'',(o'')) \end{aligned}$$

6 BIMODAL SPEECH RECOGNITION

6.1 Previous modelization

We have at our disposal two streams produced by two sources of different origins (acoustical and visual) but corresponding to the same underlying sequence of recognition units. The lack of data implied by the heaviness of the visual data acquisition system [8] have led us to suppose the statistical independence of the two sources, that lead to the use of diagonal covariance matrix in our models [2]. Each model has six emitting states, one initial and one final non-emitting state. The associated distributions are two-gaussian mixtures. Acoustic models emit vectors composed of twelve MFCC coefficients and signal energy to which we add first order derivatives. Labial models emit vectors composed of height, width, inner-lips area and first and second order derivatives. Acoustico-labial models emits the concatenation of acoustic and labial vectors described above. It is well worth noting that the two information sources are weighted, in order to give more importance to the acoustic one. The new acoustico-labial models are built by doing the product of the acoustic and labial models above, after disjoint training.

6.2 Isolated words

If two sequences of vectors represent the pronunciation of only one word of the vocabulary W and considering that an acoustic model A_i and a labial one A'_i are associated to each word w_i of W, we should calculate :

$$\arg \max_i P(A_i, A'_i | O, O')$$

Then,

$$\begin{aligned} P(A_i, A'_i | O, O') &= \frac{P(O, O', A_i, A'_i)}{P(O, O')} \\ &= \frac{P(O, O' | A_i, A'_i) \cdot P(A_i, A'_i)}{P(O, O')} \end{aligned}$$

An approximation, related to the hypothesis of statistical independence of the two sources can be done :

$$\begin{aligned} P(O, O' | A_i, A'_i) &\simeq P(O | A_i, A'_i) \cdot P(O' | A_i, A'_i) \\ &\simeq P(O | A_i) \cdot P(O' | A'_i) \end{aligned}$$

So :

$$P(A_i, A'_i | O, O') = \frac{P(O | A_i) \cdot P(O' | A'_i) \cdot P(A_i, A'_i)}{P(O, O')}$$

Then, A_i et A'_i being associated to the same word w_i of W, we have $P(A_i) = P(A'_i) = P(w_i)$. In addition, all w_i are supposed equiprobable, considering the absence of linguistic model, so $P(w_i)$ (as well as $P(O, O')$) is constant in the calculation of $\arg \max_i$:

$$\arg \max_i P(A_i, A'_i | O, O') = \arg \max_i P(O | A_i) \cdot P(O | A'_i)$$

So, we can calculate even $\arg \max_i P(O | A_i) \cdot P(O' | A'_i)$, either $\arg \max_i P''(O, O' | A''_i)$ with $A''_i = A_i * A'_i$.

6.3 Connected words

The two sequences of vectors represent the pronunciation of a sequence (of unknown length) of words w_i of the vocabulary W. An acoustic model A_i and a labial model A'_i are associated to each of these words. As a result, an acoustic model A_s and a labial one A'_s, built by concatenation of word-models, are associated to each sequence s of words. The calculation of $\arg \max_s P(O | A_s) \cdot P(O' | A'_s)$ is too long, due to the great number of possible sequences of words. So, a method could be to create the model A'' by setting all the models $A''_i = A_i * A'_i$ concurrentely linked into a loop. We can then obtain the decoding of the audio-visual message by searching the path of A'' which holds the highest probability (Viterbi algorithm) [7].

7 UNIMODAL SPEECH RECOGNITION

We can visually note the presence of asynchrony between the different formants on a sonagram. We can then feel free to think that improvement can results from a handling of these problems. Splitting the acoustic parameters into two streams, we create and train a set of models for both cesptral bands. Finally, we obtain the set of models to be used for the decoding by associating to each word the product of the model corresponding to the first cepstral band and the model corresponding to the second one.

8 EVALUATION

The corpus is composed of the pronunciation of the 26 letters of the french alphabet, continuously spoken by only one speaker. The acquisition was done by the I.C.P. laboratory of Grenoble in the AMIBE project framework [8]. The data are composed of 200 files of 4 letters, split into 70% training data and 30% test data. All the systems described above were developed and evaluated with the HTK toolkit of the C.U.E.D.

8.1 Bimodal recognition results

With the acoustic model (26 MFCC coefficients) we obtain a 87.9% letter recognition score, with the labial

one 41.3%, with the synchronous acoustico-labial one 87.5% and with the product acoustico-labial one 89.4%. The results show a 1.9% improvement as well as the importance of taking into consideration the asynchrony phenomenon between acoustic and labial data. Furthermore, this handling leads to a 1.5% gain of recognition against the purely acoustic model.

8.2 Unimodal recognition results

The MFCC coefficients of the acoustic model described above are separated into two streams. By doing so, we split the acoustic space in an arbitrary way. With these data, four sets of models are created : A set of models working on the vector composed of the first 6 MFCC coefficients and their first order derivatives (12 components), a set of models working on the vector composed of the last 6 MFCC coefficients, the signal energy and their first order derivatives (14 components), a set of models working on the complete vector (concatenation of the two vectors above), and a set of models obtained by doing the product of the first two trained models described above.

A slight difference in the performance of the acoustic models between the two experiments appears, due to a different distribution of the parameters among the streams. With the first model we obtain a 84.5% letter recognition score, with the second 73.5%, with the third 88.6% and with the last 89.8%. That shows a 1.2% improvement which confirms the hypothesis previously formulated. Results are recapitulated in table 1 below.

A	L	SB	PB
87.9	41.3	87.5	89.4
A1	A2	A3	PA
84.5	73.5	88.6	89.8

Table 1: Results, A : Acoustic, L : Labial, SB : Synchronous bimodal, PB : Product bimodal, A1 : First parameters acoustic, A2 : Last parameters acoustic, A3 : 26 parameters acoustic, PA : Product acoustic

9 CONCLUSION

We have shown that we can benefit from the handling of anticipation and retention phenomena in a continuous density HMM based system. This gain could be obtained in the bimodal speech field as well as in the more widely studied framework of unimodal speech. Moreover, the method we have suggested allows a merging of two different topologies. We can thus choose the optimal topology for each part of the information vector and for each recognition unit before merging both. Furthermore, the learning stage being only necessary for partial models, this method does not require a great number of data and does not involve complexity problems for this phase, contrary to Master-Slave models [9]. Finally, it does not require any modification of classical algorithms and so, could be applied to any Markovian system.

10 FUTURE DIRECTIONS

In order to confirm the previous results, it should be first necessary to proceed to a greater scale experiment, using more data and more speakers. Handling these phenomena outside the recognition units should also be a source of improvement. To this purpose, it will be necessary to build a specific decoding strategy.

References

[1] J. Robert-Ribes (1995)
Modèles d'intégration audio-visuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique de voyelles
Thèse de doctorat - INPG Grenoble

[2] P. Jourlin, M. El-Bèze, H. Méloni (1995)
Integrating visual and acoustic informations in a speech recognition system based on HMM
International Congress of Phonetic Sciences, Stockholm, vol. 4, 288-291

[3] C. Abry, M.T. Lallouache (1994)
Pour un modèle d'anticipation dépendant du locuteur - Données sur l'arrondissement en français
Bulletin de la communication parlée - ICP Grenoble

[4] R. André-Obrecht, B. Jacob, C. Sénac (1996)
Words on Lips : How Merging Acoustic and Articulatory Informations to Automatic Speech Recognition
European Signal Processing Conference, Trieste

[5] P. Deléglise, A. Foucault, M. Alissali (1996)
Asynchronous Integration of Audio and Visual Sources in Bi-modal Automatic Speech Recognition European Signal Processing Conference, Trieste

[6] C. Montacié, M.J. Caraty, R. André Obrecht, L.J Boë, P.Deléglise, M. El-Bèze, I. Herlin, P. Jourlin, T. Lallouache, B. Leroy, H. Méloni (1995)
Applications Multimodales pour Bornes et Interfaces Multivaluées Ecole Thématique : Traitement automatique de la parole : Fondements et Perspectives, Marseille, 155-164

[7] S.J. Young, N.H. Russel, J.H.S. Thornton (1989)
Token Passing : a Simple Conceptual Model for Connected Speech Recognition Systems
Technical Report - CUED/F-INFENG/TR.38 -

[8] M.T. Lallouache (1991)
Un poste "visage-parole" couleur
Thèse de doctorat - INPG Grenoble

[9] F. Brugnara, R. De Mori, D. Giuliani, M. Omologo (1991)
A parallel HMM approach to speech recognition
Proceedings of Eurospeech '91, Gènes, 1103-1106