STOCHASTIC TRAJECTORY MODEL WITH STATE-MIXTURE FOR CONTINUOUS SPEECH RECOGNITION

Irina Illina and Yifan Gong

CRIN/CNRS & INRIA-Lorraine,
B.P. 239, F-54506 Vandœuvre-lès-Nancy, France
{illina, gong}@loria.fr

ABSTRACT

The problem of acoustic modeling for continuous speech recognition is addressed. To deal with coarticulation effects and inter-speaker variability, an extension of the Mixture Stochastic Trajectory Model (MSTM) is proposed. MSTM is a segment-based model using phonemes as speech units. In MSTM, the observations of a phoneme are modeled by a set of stochastic trajectories. The trajectories are modeled by a mixture of probability density functions (pdf) of state sequences. Each state is associated with a multivariate Gaussian density function. In this paper, we propose to replace the state single Gaussian pdf by a mixture of Gaussian pdfs (MSTM with State-Mixture, SM-MSTM). The parameters of the model are estimated under the ML criterion, using the Expectation-Maximisation (EM) algorithm. The tests of the system on a speaker-dependent continuous speech recognition task show a reduction in the word error rate by about 15% over the baseline MSTM, even for an equal number of parameters. Experiments based on a multi-speaker continuous speech recognition task do not lead to significant improvement over the baseline system.

1. INTRODUCTION

MSTM is a segment-based model using phonemes as speech units [5]. Compared to hidden Markov models, in MSTM a mixture is defined on observation sequences rather than on individual observation vectors, thus exploiting intra-segmental information. Each component trajectory is modeled by a multivariate Gaussian pdf. Some extensions of MSTM have been proposed recently. A Semi-Continuous version of MSTM is proposed in [9]. The explicit modeling of the time evolution as a first order auto-regressive process is given in [1]. Another time evolution modeling is developed in [2] by using a mixture of polynomial functions.

Because of the context influence in continuous speech, the variability is larger in the extremities of the phone segment that in the center. In order to model the contextual effects – such as coarticulation –, some authors have proposed the use of context-dependent continuous density Gaussian mixture HMM based on state clustering [10].

For a multi-speaker recognition task, the speech variability is larger at phone level than for speaker-dependent task. In order to take this variability into account, a gender-dependent acoustic model has been proposed [4]: two sets of models (male/female) are trained, and the recognizer searches for the best solution between the two sets of models. Speaker variability could also be addressed in the MSTM framework. In [7], the idea of modeling the long term variability is presented.

Since basic MSTM uses only a single Gaussian pdf per trajectory state, its flexibility to capture large speech variability may be limited. In this paper, we are interested in modeling the variability due to coarticulation effects in speaker-dependent mode and the inter-speaker variability in multi-speaker mode. We propose to model each state of a component trajectory of MSTM by mixture of multivariate Gaussian pdfs.

The organisation of the paper is as follow. In section 2, after a presentation of MSTM, the SM-MSTM is described. Section 3 describes the ML estimate of the parameters using algorithm EM. The paper ends with the validation experiments and the discussion, followed by a conclusion.

2. ACOUSTIC MODEL

2.1. MSTM

In this section, the basic principle of MSTM is presented. We use the notation $p(X)$ for the pdf of continuous X and $Pr(X)$ for the probability of discrete X. Let X be linearly re-sampled Q points of an observation vector sequence of a speech trajectory. In MSTM a trajectory is defined as a fixed length sequence of Q observation vectors in the speech specific parameter space:

$$X = \{x_1, x_2, \ldots, x_Q\}, \quad x_i \in R^D \tag{1}$$

Each sampled point is called state. In this paper, we use a trajectory to model the context independent phone. An explicit modeling of phone duration is provided in the framework of MSTM [5]. However, for notation simplicity, any references to the duration modeling will be omitted from this point forward.

The pdf of the observed trajectory X is modeled as a mixture of output of stochastic trajectory generators associated to the phoneme

symbol a:

$$p(X|a) = \sum_{k=1}^{K^a} Pr(t_k|a) p(X|t_k, a) \quad (2)$$

where K^a is the number of components in the mixture for the symbol a, $Pr(t_k|a)$ is the *a priori* probability of the trajectory component t_k, and $p(X|t_k, a)$ is the pdf of trajectory X given t_k and a. It is important to note that, as opposed to HMM, the mixture is defined on the observation sequences rather than on individual observations. The mixture organisation of trajectories in MSTM allows to model a large variability of speech segments.

The Q states of X are assumed to be independent, given the trajectory cluster t_k and the symbol a:

$$p(X|t_k, a) = \prod_{i=1}^{Q} p(x_i|t_k, a) \quad (3)$$

Each state observation x_i is modeled as a multivariate Gaussian pdf of mean m_{ik}^a and covariance matrix Σ_{ik}^a:

$$p(x_i|t_k, a) = N(x_i; m_{ik}^a, \Sigma_{ik}^a), \quad i = 1, \dots Q \quad (4)$$

2.2. SM-MSTM

To improve the modeling of coarticulation and speaker variability we propose to define $p(x_i|t_k, a)$ as a mixture of Gaussian pdfs (state-mixture):

$$p(x_i|t_k, a) = \sum_{l=1}^{\mathcal{T}_i^{ka}} p(x_i|\tau_{lik}, t_k, a) Pr(\tau_{lik}|t_k, a) \quad (5)$$

where τ_{lik} is the state cluster associated with the state i of the trajectory cluster t_k and the phoneme a. The number of mixture components $\mathcal{T}_i^{ka}$ for the state i, the trajectory component t_k and the symbol a is determined heuristically according to the amount of available training data. $Pr(\tau_{lik}|t_k, a)$ is the *a priori* probability of state component τ_{lik}, given t_k and a. $p(x_i|\tau_{lik}, t_k, a)$ is the pdf of state x_i given τ_{lik}, t_k and a.

We assume:

$$p(x_i|\tau_{lik}, t_k, a) = N(x_i; m_{lik}^a, \Sigma_{lik}^a) \quad (6)$$

Figure 1 illustrates a SM-MSTM with one trajectory cluster ($K^a = 1$) and two clusters in each state ($\mathcal{T}_i^{ka} = 2$).

In summary, we obtain the following expression:

$$p(X|a) = \sum_{k=1}^{K^a} [Pr(t_k|a) \times \prod_{i=1}^{Q} \sum_{l=1}^{\mathcal{T}_i^{ka}} p(x_i|\tau_{lik}, t_k, a)\ Pr(\tau_{lik}|t_k, a)] \quad (7)$$

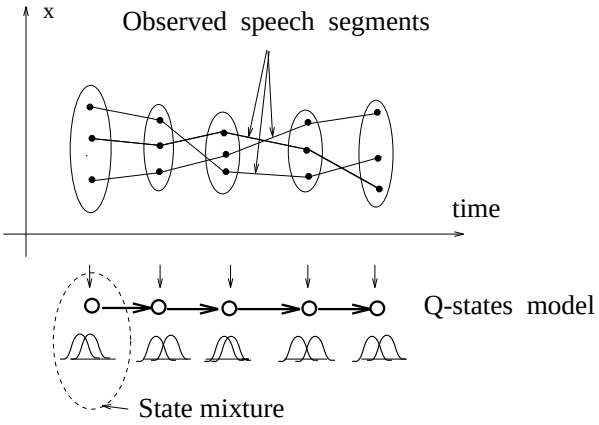

Figure 1: MSTM with State-Mixture for one symbol ($Q = 5, K^a = 1, \mathcal{T}_i^{ka} = 2$)

3. PARAMETER ESTIMATION

3.1. EM formulation

The EM algorithm is an iterative procedure for ML estimation of parameters with incomplete data. EM maximises the log-likelihood of observable data, by iteratively maximizing the expectation of the log-likelihood of the complete data (observable and unobservable). The observable data consists of measured sequence **X**. The unobservable data **Y** contains some hidden information. The EM algorithm uses the following expression:

$$\max_{\lambda} Q(\lambda|\lambda') = \max_{\lambda} E\left\{\log p(\mathbf{X},\mathbf{Y}|\lambda) | \mathbf{X}, \lambda'\right\} \quad (8)$$

where λ denotes the parameter set. The n^{th} iteration of the EM algorithm consists in two distinct steps:

$$\text{E-step : compute } Q(\lambda|\lambda^n)$$
$$\text{M-step : choose } \lambda^{n+1} = \arg\max_{\lambda} Q(\lambda|\lambda^n)$$

An important property of the EM algorithm is that $\log p(\mathbf{X}|\lambda')$ is guaranteed to improve monotonically until it reaches a stationary point. It is also known that the sequence $\{\lambda^n\}$ converges [3]. The defect of EM algorithm is that it finds only a local maximum of $Q(\lambda|\lambda^n)$.

3.2. SM-MSTM parameter derivation

Let the observable data be:

$$\mathbf{X} = \{\{X = \{x_i\}\}^a\}, \quad i = 1, \dots, Q, \quad a \in \mathbb{P}$$

and the associated unobservable data be:

$$\mathbf{Y} = \{\{t_k, \tau_{lik}\}^a\},$$
$$l = 1, \dots \mathcal{T}_i^{ka}, i = 1, \dots Q, \quad k = 1, \dots K^a, \quad a \in \mathbb{P}$$

where t_k is an unobservable integer between 1 and K^a, indicating the number of trajectory clusters associated to the symbol a, and τ_{lik} is an unobservable integer between 1 and $\mathcal{T}_i^{ka}$, indicating the

number of the state clusters associated to the state i, the trajectory t_k and the symbol a. $\mathbb{P}$ is the set of phonetic symbols. Therefore:

$$Q(\lambda|\lambda') = E\left\{\log p(\{\{(X, t_k, \tau_{l1k}, \ldots, \tau_{lQk},)\}^a\}|\lambda)|\{\{X\}^a\}, \lambda'\right\} \quad (9)$$

For notation simplicity and without loss of generality, in the remaining of this paper we will limit ourselves to one symbol. Using equations of section 2 and denoting $\zeta_{lik} \triangleq Pr(\tau_{lik}|t_k, \lambda)$ and $\eta_k \triangleq Pr(t_k|\lambda)$, Eq-9 can be expressed as:

$$Q(\lambda|\lambda') = \sum_{X\in\mathcal{X}} \sum_{k=1}^{K} \sum_{i=1}^{Q} \sum_{l=1}^{\mathcal{T}_i^k} Pr(t_k, \tau_{lik}|X, \lambda') \times [\log N(x_i; m_{lik}, \Sigma_{lik}) + \log \zeta_{lik} + \log \eta_k] \quad (10)$$

where $\mathcal{X}$ is the set of all observed X. ζ_{lik} and η_k satisfy:

$$\sum_{k=1}^{K} \eta_k = 1, \quad \sum_{l=1}^{\mathcal{T}_i^k} \zeta_{lik} = 1, \quad i = 1, \ldots Q, \quad k = 1, \ldots K \quad (11)$$

The parameter set for EM-estimation can be written as:

$$\lambda \triangleq \{\eta_k, \zeta_{lik}, m_{lik}, \Sigma_{lik}\}, \quad l = 1, \ldots \mathcal{T}_i^k, \quad i = 1, \ldots Q, \quad k = 1, \ldots K \quad (12)$$

We give below the EM-reestimate for each parameter:

- *a priori* probability of trajectory:

$$\eta_k = \frac{\sum_{X\in\mathcal{X}} Pr(t_k|X, \lambda')}{|\mathcal{X}|} \quad (13)$$

  where $|\mathcal{X}|$ stands for the cardinal of the set $\mathcal{X}$.

- *a priori* probability of state:

$$\zeta_{lik} = \frac{A_{lik}}{\sum_{l=1}^{\mathcal{T}_i^k} A_{lik}} \quad (14)$$

  where $A_{lik} \triangleq \sum_{X\in\mathcal{X}} Pr(\tau_{lik}, t_k|X, \lambda')$

- mean vector and covariance matrix:

$$m_{lik} = \frac{\sum_{X\in\mathcal{X}} B_{lik} x_i}{A_{lik}} \quad (15)$$

$$\Sigma_{lik} = \frac{\sum_{x\in\mathcal{X}} B_{lik} \left(x_i - m_{lik}^{(x_i)}\right) \left(x_i - m_{lik}^{(x_i)}\right)^{\#}}{A_{lik}} \quad (16)$$

  where $\#$ stands for transposition operation and

$$B_{lik} \triangleq Pr(\tau_{lik}, t_k|X, \lambda') \quad (17)$$

Description	CEA		AGIMMO	
	Train	Test	Train	Test
Nbr of speakers	4	4	10	10
Sentences/speaker	79	241	140	160
Gram. vocabulary	–	2010	–	1010
Word-pair perplexity	–	48	–	29
SNR (dB)	–	–	15	15

Table 1: Summary of CEA and AGIMMO corpora

$Pr(\tau_{lik}, t_k|X, \lambda')$ is obtained by:

$$Pr(\tau_{lik}, t_k|X, \lambda') = \frac{p(x_i|\tau_{lik}, t_k, \lambda') Pr(\tau_{lik}|t_k, \lambda') Pr(t_k|\lambda')}{P(X|\lambda')} \times \prod_{j=1,\ldots Q, j\neq i}^{Q} p(x_j|t_k, \lambda') \quad (18)$$

The initialisation of parameter set for EM algorithm can be performed using any adequate unsupervised classification technique. In our experiments, the LBG algorithm is used [8].

4. EXPERIMENTS AND RESULTS

4.1. Task description

Experiments deal with two French continuous speech corpora recorded by the CRIN laboratory: CEA and AGIMMO. Table 1 describes these corpora. Speech signal is sampled at 16 kHz. 13^{th} order mel-cepstral vectors were computed every 10 ms with an analysis window of 32 ms. In average, there are about 70 observations per phoneme for each speaker in the training part of CEA corpus and about 80 observations in the AGIMMO corpus. For AGIMMO corpus, among 160 testing-sentences, 60 are common to all speakers and the remaining 100 are speaker specific.

The two tasks are difficult because of insufficient training data, noisy recording conditions (AGIMMO), and because the pauses between word are not modeled by our grammar.

4.2. Experiment design and discussion

We tested the system on two tasks. The first is a speaker-dependent continuous speech recognition task with CEA corpus. The second is a multi-speaker continuous speech recognition task with AGIMMO corpus. In this last task, an unique system is trained with the data from all speakers. For each corpus, 32 context-independent phone models, including one silence model, are built. The number of state Q is 5. The number of trajectory mixtures ($|K^a|$) and the number of state mixtures ($|\mathcal{T}_i^{ka}|$) are proportional to the number of observations in the training data. In the baseline MSTS configuration, up to 8 ($|K^a| = 8$) clusters of trajectories are used in the CEA corpus and up to 16 in the AGIMMO corpus. In the proposed SM-STM, we use only up to 4 ($|K^a| = 4$) clusters of trajectories in the CEA corpus and up to 8 in the AGIMMO corpus and 2 ($|\mathcal{T}_i^{ka}| = 2$) state clusters.

speaker	MSTM	SM-MSTM
alv	97.23	97.77
flf	97.91	98.85
pab	98.45	98.04
yfg	97.37	97.64
AVG	97.74	98.08

Table 2: Word accuracy rates as function of speakers and models for CEA corpus

speaker	MSTM	SM-MSTM
brs	87.96	89.57
crm	85.65	86.56
jel	72.15	74.50
lar	96.06	91.97
ols	95.60	94.78
sat	80.69	80.36
std	90.77	86.00
syc	85.60	87.68
vil	92.41	93.70
yig	90.99	93.89
AVG	88.00	88.08

Table 3: Word accuracy rates as function of speakers and models for AGIMMO corpus

The speaker-dependent task was tested in the following configuration: state-mixture (Eq-5) for the first and the last states and a simple Gaussian (Eq-4) for remaining states. This configuration emphasises modeling of the variability due to phonetic contexts. The result of the test is given in Table 2, which shows that the performance of the proposed model is better than MSTM, even for an equal number of parameters. We observe that the MSTM system gives 97.74% word accuracy (21 Del., 105 Sub. and 8 Ins. over 5928 words), and the SM-MSTM gives 98.08% word accuracy (14 Del., 92 Sub. and 8 Ins.), which represents about 15% of error reduction. These experiments suggest that the MTSM-SM is able to take into account the variability due to phonetic contexts.

For the multi-speaker task, the state-mixture is created for each state to model the inter-speaker variability. The results are presented in Table 3. The MSTM system gives 88.0% word accuracy (58 Del., 550 Sub. and 162 Ins. over 6418 words). The SM-MSTM gives 88.08% word accuracy (51 Del., 552 Sub. and 162 Ins.). This configuration does not give any significant improvement. We may consider that it is due to a trajectory folding phenomenon [6]: clusters of trajectories cannot be well represented, because the information on the continuity of each individual trajectory is lost. Experiments to validate this hypothesis are in progress.

5. CONCLUSION

In this paper, the Mixture Stochastic Trajectory Model with State-Mixture is proposed. The main difference between SM-MSTM and MSTM lies in the modeling of each state pdf by a mixture of Gaussian pdf in SM-MSTM in order to increase the capability of MSTM to model the large phonetic contexts and the inter-speaker variability. We propose an ML estimate of the parameters of SM-MSTM using the EM algorithm. Experiments indicate that the SM-MSTM is able to take into account the variability due to phonetic contexts. However, our tests do not suggest significant improvement when attempting to model the inter-speaker variability. One explanation of this fact could be that the introduction of state mixture distributions increases the trajectory folding. It is plausible that the suitable balance between the number of trajectory mixtures and the number of state mixtures would give better result. Experiments to validate this hypothesis are in progress.

6. REFERENCES

1. M. Afify, Y. Gong, and J. P. Haton. Stochastic trajectory model for speech recognition: an extension to modeling time correlation. *In Proc. of European Conference on Speech Communication and Technology*, pages 515–518, September 1995. Madrid, Spain.
2. C. Cerisara, Y. Gong, and J. P. Haton. Reconaissance de la parole continue par le modèle STM polynômial. *In Actes des 21-émes Journées d'Études sur la Parole*, 1996. Avignon.
3. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.
4. J. L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker. The LIMSI continuous speech dictation system: Evaluation on the ARPA Wall Street Journal Task. *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1:557–560, April 1994. Adelaide, Australia.
5. Y. Gong and J. P. Haton. Stochastic trajectory modeling for speech recognition. *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1:57–60, April 1994. Adelaide, Australia.
6. Y. Gong, J. P. Haton, and J. F. Mari. *Issues in acoustic modeling of speech for automatic speech recognition*. INFIX Sankt Augustin, September 1994.
7. Y. Gong, I. Illina, and J. P Haton. Modeling long term variability information in mixture stochastic trajectory framework. *In Proc. of Int. Conf on Spoken Language Processing*, October 1996. Philadelphia, PA, USA.
8. Y. Linde, A. Buzo, and R. M. Gray. An algorithm for the vector quantizer design. *IEEE Trans. on Communication*, COM-28(1):84–95, January 1980.
9. O. Siohan and Y. Gong. A semi-continuous stochastic trajectory model for phoneme-based continuous speech recognition. *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, May 1996. Atlanta, Georgia, USA.
10. S. J. Young and P. C. Woodland. State clustering in hidden Markov model-based continuous speech recognition. *Computer Speech and Language*, 8:369–383, 1994.