ELIMINATION OF TRAJECTORY FOLDING PHENOMENON: HMM, TRAJECTORY MIXTURE HMM AND MIXTURE STOCHASTIC TRAJECTORY MODEL

Irina Illina

CRIN/CNRS, INRIA-Lorraine
54506 Vandœuvre-lès-Nancy, France
illina@loria.fr

Yifan Gong

Speech Research
Personal Systems Laboratory
Texas Instruments
Dallas TX 75265, U.S.A.
ygong@schubert.csc.ti.com

ABSTRACT

In this paper, a study of topology of Hidden Markov Model (HMM) used in speech recognition is addressed. Our main contribution is the introduction of the notion of trajectory folding phenomenon of HMM. In complex phonetic contexts and in speaker-variability, this phenomenon degrades the discriminability of HMM. The goal of this paper is to give some explanation and experimental evidence suggesting the existence of this phenomenon. The systems eliminating (partially or entirely) the trajectory folding are HMM with a special topology, called Trajectory Mixture HMM (TMHMM), and a Mixture Stochastic Trajectory Model (MSTM), proposed recently. HMM, TMHMM and MSTM have been tested on a 1011 words vocabulary, speaker dependent and multi-speaker continuous French speech recognition task. With similar number of model parameters, TMHMM and MSTM cuts down the error rate produced by the HMM, which confirms our hypothesis.

1. INTRODUCTION

Our purpose is to study the ability of HMM topology to model the structure of speech acoustic variability, due to complex phonetic context or inter-speaker variability. It is well know that the increase in the variability of speech decreases the discriminability of HMM. To improve the discriminability of HMM, different solutions are possible: complex models with high number of observation pdfs, context dependent phoneme models, dynamic coefficients or other topology of the model.

We consider that the classic topology of HMM (see Figure-1a for example) may not be adequate to modeling the structure of large speech data, and that such HMM mixes up the different sources of speech variability therefore has limited discriminability. Our hypothesis is that this is due to trajectory folding phenomenon. The goal of our work is to give some explanation and experimental evidence suggesting the existence of this phenomenon in HMM.

The organisation of the paper is as follows. In section 2, the trajectory folding phenomenon is illustrated through an example. We present also in this section some recent models and how trajectory folding is treated in these models. Section 3 provides the presentation of the continuous speech recognition task and the set of experiments. The paper ends with a summary of the results and a conclusion.

2. TRAJECTORY FOLDING PHENOMENON

The main purpose of this section is to point out trajectory folding phenomenon. To describe this phenomenon, let us take an example. Consider a 2 state (A,B), left-to-right, no skip, continuous density mixture HMM with two normal pdfs per state. Assume that, for a sound s, the model is trained with observations from 2 speakers, one male (M) and one female (F). There is strong reason to believe that, at each state, one of the pdfs will model the male voice and the other female voice.

Let us denote the pdf of the mixture component k at state j by $f_{k,j}(x)$ ($k \in \{M, F\}$ and $j \in \{A, B\}$). Obviously, from the training data, two trajectories should be identified. They can be denoted by the pdf with highest output: $f_{M,A}(x) \rightarrow f_{M,B}(x)$ for the male speaker, and $f_{F,A}(x) \rightarrow f_{F,B}(x)$ for the female speaker.

However, when recognizing an utterance, the HMM cannot prevent the model from giving a high output via the trajectories $f_{M,A}(x) \rightarrow f_{F,B}(x)$ or $f_{F,A}(x) \rightarrow f_{M,B}(x)$, which have never been observed in the training set of s. Unfortunately, such trajectories may correspond to another sound $s' \neq s$. As consequence, s' could be misrecognized as s. In other words, clusters of speech trajectories cannot be well represented, because the information on the continuity of each trajectory is lost and trajectories are folded state-by-state. We call this phenomenon **trajectory folding**.

Folding can happen along any sequence of states and a self-transition can also cause the trajectory folding. If an HMM is trained with multi-speaker data or in complex phonetic contexts, this phenomenon decreases the discriminability of HMM.

One solution to this problem is to change the topology of the model. In the following, we discuss some recent works concerning the topology of the model and relate them to our work. We show also how a model, proposed in each of these works, solves the problem of trajectory folding.

One alternative of HMM, called Structured Markov Model (SMM), has been proposed in [1]. To improve the discriminability of HMM, a solution is to increase the number of components (clusters) of state mixtures. As a consequence, HMM implicitly allows more and more combinations of mixture components along a state sequence which are never observed in the training data. This leads to saturation effects in the recognition rate. In order to reduce this and to model the structural aspects of the speech variability, SMM models the structure of the speech units as a graph of normally distributed acoustic events and a transition network (see Figure 1b for topology example). The maximum-likelihood training of SMM adjusts the transition probabilities and provides explicit transition descriptions between the state clusters instead of implicit description in HMM. A speaker-independent speech recognition evaluation shows the superiority of SMM compared to state-mixture HMM. This study shows, that in the case of high number of state clusters, classic HMM is not adequate for modeling the structure of state transitions.

In SMM, an explicit description of the transitions between the state clusters and a deletion of the transitions with zero value of the estimated probability, eliminate to some extent trajectory folding.

To cope with the problem of speaker time-drifting and inhomogeneous data sources, a Trajectory Mixture HMM has been proposed in [2]. The authors give two motivations to use TMHMM. First, for inhomogeneous data sources, HMM gather indifferently the variability from different sources, which leads to lower modeling accuracy. Second, some of the variability, or rather the inhomogeneity of data, may be known *a priori* and modeled separately. The TMHMM is considered as a large HMM with multiple trajectories as mixtures (see Figure 1c for topology example). Experiments in speaker recognition, using the syllable-like phonetic units, confirm the efficiency of TMHMM for modeling the variability due to inhomogeneous data sources. This model shows, that in the case of high variability in the data, HMM is not able to discriminate different data sources and to provide a model with good discriminability.

In TMHMM, the creation of several HMMs to represent the different speech trajectories, instead of one state-mixture HMM, partly avoids trajectory folding. The pdf of an observation sequence X, given the acoustic model λ, is:

$$p(X|\lambda) = \sum_{\mu_k \in M} Pr(\mu_k) p(X|\mu_k, \lambda) \qquad (1)$$

where μ_k denotes the k-th HMM cluster (defined on a state sequence) and $Pr(\mu_k)$ the corresponding mixture weight. The expression 1 shows that the trajectory folding is partially eliminated in this model.

An other alternative to HMM is a segment-based Mixture Stochastic Trajectory Model (MSTM), using phonemes as speech units [3, 4].

The basic idea of MSTM is the following: speech is considered as a point that moves in the parameter space as the articulatory system changes. A sequence of moving points is called a trajectory of speech. A trajectory takes into account not only the geometric position of speech observations, but also the curve of articulatory moving. For detailed representation of speech variability and high discriminability in complex acoustic contexts, trajectories are organised into clusters. As opposed to HMM, in MSTM, there are two levels of mixture pdfs. The first level of mixture is defined on observation sequences rather than on observation frames (see Figure 1d for topology example). The second one can be defined on each individual states (not shown in Figure 1d) [5]. Compared to SMM, MSTM also models explicitly the transitions. As TMHMM, MSTM models different data sources separately by different trajectory clusters. Experiments in speaker dependent speech recognition show the efficiency of this model.

To be more specific, in MSTM, in order to model the clusters of trajectories, the corresponding trajectory is represented by mixture of pdfs of fixed-length observation vector sequence X:

$$\begin{aligned} X &= \{x_1, \ldots, x_Q\} \\ p(X|\lambda) &= \sum_{t_k \in T} Pr(t_k|\lambda) p(X|t_k, \lambda) \\ &= \sum_{t_k \in T} Pr(t_k|\lambda) \prod_{i=1}^{Q} p(x_i|t_k, \lambda) \end{aligned} \qquad (2)$$

where $Pr(t_k|\lambda)$ is the probability of trajectory cluster t_k given the model λ, and $p(X|t_k, \lambda)$ denotes the pdf of X given t_k and λ. In order to model durational constraints, the time line of the observation is rescaled. The rescaling consists in mapping a segment of observed trajectory into fixed prespecified Q points of the trajectory models. The expression 2 shows that the trajectory folding is completely eliminated in this model.

3. EXPERIMENTS AND RESULTS

In order to confirm our hypothesis about the trajectory folding phenomenon and to measure the impact of this phe-

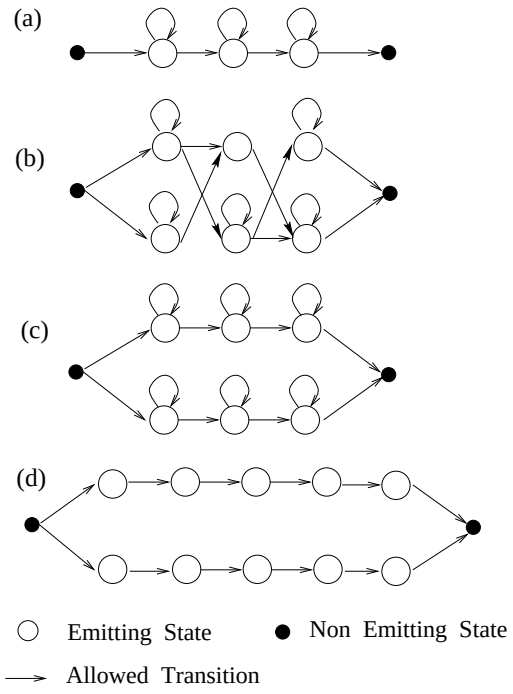

Figure 1: Topology example of models. (a) HMM (3 emitting states); (b) SMM (6 emitting states); (c) TMHMM (2 trajectories, 6 emitting states); (d) MSTM (2 trajectories, 10 emitting states)

nomenon, we have compared the recognition results obtained with HMM, TMHMM and MSTM on a same speech recognition task.

3.1. Database description

Experiments deal with a French continuous speech corpus recorded by our laboratory and corresponding to an application with CEA, the national nuclear energy agency. For training, 79 phonetically rich sentences were read by 7 French speakers (1 female). In average, there are about 70 observations per phoneme for each speaker. For testing, 241 sentences were recorded. There is only a small overlap between training and test vocabularies. Speech is sampled at 16 kHz. The observation vectors are 14 MFCC including a normalized energy computed every 10 ms with an analysis window of 32 ms. For this corpus, 32 context-independent phone models, including one silence model, are built. The language model has a word-pair equivalent perplexity of 31 and a 1013 words vocabulary. In all experiments, the covariance matrix is assumed to be diagonal. The task is difficult because of insufficient amount of training data and because of between word pauses which are not modeled by our grammar.

spkr	HMM		TMHMM		MSTM	
	%Acc	D,S,I	%Acc	D,S,I	%Acc	D,S,I
alv	97.71	9,25,0	98.18	6,21,0	99.19	0,10,2
dof	97.30	7,27,6	98.65	3,17,0	98.72	0,18,1
fff	97.84	8,19,5	97.98	5,23,2	99.53	1,4,2
loc	97.17	13,28,1	97.50	6,29,2	99.39	0,8,1
ols	99.26	3,8,0	99.12	2,9,2	99.60	0,5,1
pab	98.58	4,15,2	98.99	1,12,2	99.66	0,5,0
yfg	97.57	6,30,0	98.31	2,23,0	99.06	2,11,1
AVG	**97.91**	**50,152,14**	**98.38**	**25,134,8**	**99.31**	**3,61,8**

Table 1: Word accuracy rates as function of speakers and models for CEA corpus in speaker dependent mode. $\%Acc$ - $\%$ Accuracy, D - deletions, S - substitutions, I - insertions.

3.2. Recognizers and experiment design

For the experiments, we have developed HMM and TMHMM using HTK V1.5 [6]. HTK, developed by the Speech Group at Cambridge University Engineering Department, is a software toolkit for building and manipulating HMM systems. We have tested each system in two configurations: in speaker dependent mode (SD), and in multi-speaker mode (MS), where we train a system with the data from all speakers. The all tests, we tied to keep the total number of pdfs approximately equal for HMM, TMHMM and MSTM for each configuraion. The HMM used is a 3 states, left-to-right, no skip model (Figure 1a). In the HMM, the number of mixture components (normal distributions) per state is 2 (larger number decreases recognition accuracy) for SD configuration and $2, 4, 8$ for MS configuration. In the TMHMM, this number is 1 for SD configuration and 1 and 2 for MS configuration. For TMHMM and MSTM, we use 2 trajectory components (normal distributions) ($|T| = 2, |M| = 2$) for SD (Figure 1c, 1d) , and $2, 4, 8$ components for MS. In MSTM, the number of states is 5 ($Q = 5$). For the initialisation of trajectory component of TMHMM, we use trajectories given by MSTM. The HMM and TMHMM use 10 iterations of Baum-Welch estimation and 4 cycles of embedded reestimation.

3.3. Summary of the results

Results in terms of word recognition accuracy of continuous speech recognition in speaker dependent mode, are given in Table 1. The HMM system gives 97.91% word accuracy (with 50 deletions, 152 substitutions and 14 insertions over 10374 words, the 95% confidence interval is $97.6\% - 98.2\%$). We observe that TMHMM gives higher recognition rate as expected (98.38% word accuracy with 25 del., 134 sub. and 8 ins. over 10374 words , the 95% confidence interval is $98.1\% - 98.6\%$). This represents reduction of 22% of the error rate produced by the HMM system. The TMHMM

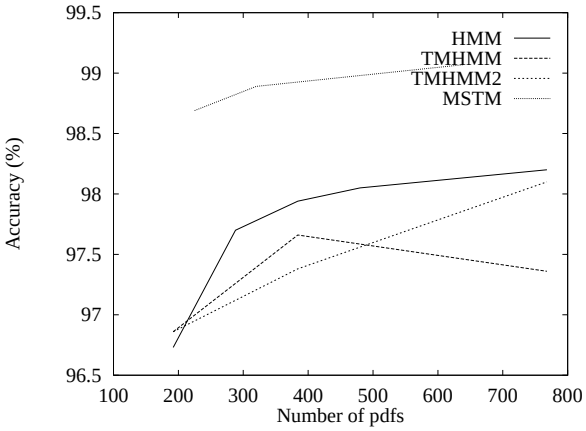

Figure 2: Word accuracy rates as function of number of pdfs and models for CEA corpus in multi-speaker mode

have a smaller insertion and deletion rate than HMM. The highest recognition rate is obtained using MSTM (99.31% word accuracy , with 3 del., 61 sub. and 8 ins. over 10374 words, a 95% confidence interval is $99.1\% - 99.4\%$). This model have a same insertion rate and substantially smaller deletion and substitution rate than TMHMM. These results show that in TMHMM the decreasing of the number of implicitelly modeled speech trajectories gives the better accuracy compared to HMM and that, in this case, two explicit trajectories of MSTM are enough to give the highest recognition rate. This fact confirms our hypotheses of trajectory folding of HMM due to complex phonetic context.

Results for multi-speaker mode are presented in Figure 2. In this task, each system is trained with the data from all speakers. The word accuracy of different models is compared over the total number of pdfs. The comparaison shows that MSTM gives the highest recognition rate. TMHMM with one Gaussian pdf per state (TMHMM in the Figure 2) and TMHMM with two Gaussian pdf per state (TMHMM2 in the Figure 2), show lower accuracy than HMM, which we attribute to the insufficient amount of training data. In the TMHMM the training data are divided in $|M|$ trajectory clusters, which may not yield reliable trajectory parameter estimation.

4. CONCLUSION

We introduced in this paper the notion of trajectory folding phenomenon of Hidden Markov Models (HMM): clusters of speech trajectories cannot be well represented, because the information on the continuity of each trajectory is lost and trajectories are folded. This phenomenon degrades the discriminability of HMM in complex phonetic contexts and in large speaker and speech variability. Our claim is that this phenomenon can be partially avoided in Trajectory Mixture HMM (TMHMM) and completely avoided in the Mixture Stochastic Trajectory Model (MSTM). We have described how trajectory folding is dealt with in HMM, TMHMM and MSTM. Experiments with HMM, TMHMM and MSTM on a continuous speech recognition task in speaker dependent mode confirm our hypothesis.

5. REFERENCES

[1] F. Wolfertstetter and G. Ruske. Structured Markov models for speech recognition. *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1:544–547, April 1995.

[2] J. Su, H Li, J. P. Haton, and K. T. Ng. Speaker time-drifting adaptation using trajectory mixture hidden Markov models. *In Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2:709–712, May 1996. Atlanta, Georgia, USA.

[3] 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.

[4] Y. Gong. Stochastic trajectory modeling and sentence searching for continuous speech recognition. *IEEE Trans. on Speech and Audio Processing*, January 1997.

[5] I. Illina and Y. Gong. Stochastic trajectory model with state-mixture for continuous speech recognition. *In Proc. of Int. Conf on Spoken Language Processing*, 1, October 1996. Philadelphia, PA, USA.

[6] S. J. Young, P. C. Woodland, and W. J. Byrne. *HTK Version 1.5: User, Reference and Programmer Manual.* Cambridge University Engineering Department and Entropic Research Laboratories Inc., September 1993.