

MODELING LONG TERM VARIABILITY INFORMATION IN MIXTURE STOCHASTIC TRAJECTORY FRAMEWORK

Yifan Gong, Irina Illina and Jean-Paul Haton

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

ABSTRACT

The problem of acoustic modeling for speech recognizers is addressed. We distinguish two types of speech variability, long term (speaker identity, stationary noise, channel distortion) and short term (phoneme class). Currently, most recognizers model the two variabilities without considering their specificities, which may result in flat distributions with limited discriminability. In our system, the long term variability (*environment*) is modeled by a mixture model, where each mixture is modeled by a Mixture Stochastic Trajectory Model (MSTM). We propose the Environment Dependent Mixture Stochastic Trajectory Model (ED-MSTM) to model a set of environments. The parameters of ED-MSTM are estimated using the Maximum Likelihood (ML) estimation criterion by the Expectation-Maximisation (EM) algorithm. Our model has been tested on a 1011 word vocabulary, multi-speaker continuous French recognition task with noisy speech. In the experiments, we assume that speakers can be grouped into a pre-determined number of classes and that class label of a speaker is missing. The use of environmental modeling cut down the error rate produced by the multi-speaker system by about 15%, which is a statistically significant improvement. The idea of environment modeling is applicable to other acoustic modeling techniques such as Hidden Markov Models.

1. INTRODUCTION

We categorise speech variabilities into two types: long term variability and short term variability. Long term variability corresponds to speaker identity, stationary background noise or channel distortion. Such variabilities do not carry information about the vocal message, and should therefore be eliminated during recognition. On the other hand, short term variability corresponds to articulation changes and should be recognised. The distinction between the two variabilities allows the modeling of long term changes within the entire utterance and short term changes across different acoustic models.

In many recognizers, the two types of variabilities are usually mixed up and represented by a same acoustic modeling scheme, which may limit the recognition performance. In this paper, we identify long term variability as *environment*. We extend the mixture stochastic trajectory models (MSTM) to cope with environmental

changes.

MSTM models a speech segment by a mixture of trajectories, each of which is a sequence of multi-dimensional probability density functions (pdfs) [6]. Phoneme duration distributions are also included. A Semi-Continuous version of MSTM is presented in [9]. The adaptation of the topology of MSTM to deal with the coarticulation effects and inter-speaker variability is proposed in [7]. The explicit modeling of the time evolution as a first order autoregressive process is given in [1]. Time evolution modeling by mixture of polynomial functions for MSTM is presented in [2].

In this paper, the basic assumption is that a long term variable can be observed and is generated by a finite number of random sources, each with an *a priori* probability. However, the information about which source actually generated the observation is missing, which requires that the environment be stochastically modeled as a hidden variable.

The organisation of the paper is as follows. In section 2, we present the Environment Dependent model and its differences with our baseline MSTM. The section 3 describes the ML estimate of the parameters using EM-algorithm. The validation experiments and the conclusion are given in section 4 and 5.

2. ENVIRONMENT DEPENDENT ACOUSTIC MODELING

In MSTM, the observation of a speech segment X is assumed to be generated by a set of random segmental trajectory generators, each with an *a priori* probability [6]. Compared to hidden Markov models, in MSTM a mixture is defined on observation sequences (component trajectories) rather than on individual observations, thus exploiting intra-segmental information.

In this paper, we denote $p(X)$ the continuous pdf of X and $Pr(X)$ the probability of discrete X .

We introduce two kinds of observations: acoustic observation vector sequence X , and an utterance-specific environment measurement y . Let \mathbb{P} be the set of phonetic symbols, \mathcal{X}^a the set of trajectories taken from the training corpus, associated to the symbol $a \in \mathbb{P}$. Let $p(X|y, a, d)$ be the pdf of an observed trajectory $X \in \mathcal{X}^a$, given

the environment observation $y \in \mathcal{Y}$, the phoneme symbol $a \in \mathbb{P}$ and the duration d . \mathcal{Y} is the set of environment measurements. In this section, we extend the pdf $p(X|a, d)$ of MSTM to $p(X|y, a, d)$ of ED-MSTM. Let $e \in \mathcal{E}$ be the environment source that generated y and $t \in \mathcal{T}^a$ be the trajectory source generated X , where \mathcal{E} is the set of environments and \mathcal{T}^a is the set of trajectories associated to the symbol a . Both e and t are unobservable. We assume that e and t are discrete. The joint pdf of variables X, y, e, t , given a and d , is:

$$p(X, y, e, t|a, d) = p(X|y, e, t, a, d) \times p(y|e, t, a, d)Pr(t|e, a, d)Pr(e|a, d) \quad (1)$$

In order to reduce the number of free parameters, we make four assumptions:

- the acoustic observation is independent of the environment measurement:

$$p(X|y, e, t, a, d) = p(X|e, t, a, d) \quad (2)$$

- the environment measurement is independent of trajectory, duration and phoneme:

$$p(y|e, t, a, d) = p(y|e) \quad (3)$$

- the trajectory probability is independent of duration:

$$Pr(t|e, a, d) = Pr(t|e, a) \quad (4)$$

- the environment probability is independent of duration and phoneme:

$$Pr(e|a, d) = Pr(e) \quad (5)$$

Therefore:

$$p(X, y, e, t|a, d) = p(X|e, t, a, d)p(y|e)Pr(t|e, a)Pr(e) \quad (6)$$

Normally, the duration information d of trajectory for MSTM-SM is used in a similar way as in MSTM [6]. For notation simplicity we do not use it in the following.

3. EM ESTIMATION OF PARAMETERS

We use the Maximum Likelihood criterion to find model parameters. The EM-algorithm gives an efficient solution to such a problem and guarantees the improvement of training data likelihood until a stationary point [3]. EM iteratively maximises, with respect to the new parameter set λ , the mathematical expectation of the log-likelihood of the complete data (\mathbf{X}, \mathbf{Y}) , conditioned on the observed data \mathbf{X} , and for a value λ' of the parameter. The expectation is taken over the sample space of the unobservable data \mathbf{Y} .

$$Q(\lambda|\lambda') \triangleq E \{ \log p(\mathbf{X}, \mathbf{Y}|\lambda) | \mathbf{X}, \lambda' \} \quad (7)$$

In our case:

$$\mathbf{X} = \{y, \{X\}^a | a \in \mathbb{P}\} \text{ and } \mathbf{Y} = \{e, \{t\}^a | a \in \mathbb{P}\} \quad (8)$$

where $\{X\}^a$ is the set of observed training data for the phonetic symbol a and y is the observed environment. $\{t\}^a$ is the set of trajectory sources associated to $\{X\}^a$ and e is the environment source associated to y . Therefore:

$$Q(\lambda|\lambda') = \sum_{a \in \mathbb{P}} \sum_{X \in \mathcal{X}^a} \sum_{y \in \mathcal{Y}} \sum_{e \in \mathcal{E}} \sum_{t \in \mathcal{T}^a} \log p(X, y, e, t|\lambda)p(e, t|X, y, \lambda') \quad (9)$$

where \mathcal{X}^a and \mathcal{Y} are respectively the acoustic observations of symbol a and the environment measurements of the training data.

In order to maximise $Q(\lambda|\lambda')$ (according to Eq-9) we must specify the form of elementary pdfs for X and y . We assume:

- a trajectory is sampled into Q points:

$$X \triangleq \{x_1, x_2, \dots, x_Q\}, x_i \in R^D \quad (10)$$

- each point is assumed to be an independent multivariate Gaussian pdf, given the trajectory:

$$p(X|e, t, \lambda) = \prod_{i=1}^Q N(x_i; m_{i,e,t}^{(x)a}, \Sigma_{i,e,t}^{(x)a}), \quad (11)$$

$$p(X|e, \lambda) = \sum_{t \in \mathcal{T}^a} p(X|e, t, \lambda)p(t|e, \lambda) \quad (12)$$

- the environment measurement is modeled as a multi-dimensional Gaussian pdf:

$$p(y|e, \lambda) = N(y; m_e^{(y)}, \Sigma_e^{(y)}), \quad (13)$$

$$p(y|\lambda) = \sum_{e \in \mathcal{E}} p(y|e, \lambda)Pr(e|\lambda) \quad (14)$$

For notation simplicity, we denote:

$$\zeta_{e,t}^a \triangleq Pr(t|e, \lambda) \text{ and } \eta_e \triangleq Pr(e|\lambda) \quad (15)$$

$\zeta_{e,t}^a$ and η_e satisfy stochastic constraints:

$$\sum_{e \in \mathcal{E}} \eta_e = 1 \text{ and } \forall e \in \mathcal{E}, \forall a \in \mathbb{P}, \sum_{t \in \mathcal{T}^a} \zeta_{e,t}^a = 1 \quad (16)$$

Eq-9 can now be expressed as:

$$\begin{aligned} Q(\lambda|\lambda') &= \sum_{a \in \mathbb{P}} \sum_{X \in \mathcal{X}^a} \sum_{y \in \mathcal{Y}} \sum_{e \in \mathcal{E}} \sum_{t \in \mathcal{T}^a} p(e, t|X, y, \lambda') \\ &\times \left[\sum_{i=1}^Q \log N(x_i; m_{i,e,t}^{(x)a}, \Sigma_{i,e,t}^{(x)a}) \right. \\ &\left. + \log N(y; m_e^{(y)}, \Sigma_e^{(y)}) + \log \zeta_{e,t}^a + \log \eta_e \right] \quad (17) \end{aligned}$$

with the parameter set λ :

$$\lambda \triangleq \{\lambda^a \triangleq \{m_{i,e,t}^{(x)a}, \Sigma_{i,e,t}^{(x)a}, m_e^{(y)}, \Sigma_e^{(y)}, \zeta_{e,t}^a, \eta_e\}\} \quad (18)$$

Next, we maximise $Q(\lambda|\lambda')$ with respect to each parameter in λ , under the stochastic constraints (Eq.-16). The maximisation can be achieved by using the Lagrange constrained optimisation method, which involves taking the partial derivative of the Lagrangian equation with respect to the parameters and solving the resulting equations for the parameters.

For notation simplicity, we use the following expressions:

$$A^a(e, t|\lambda') \triangleq \sum_{X \in \mathcal{X}^a} \sum_{y \in \mathcal{Y}} p(e, t|X, y, \lambda') \quad (19)$$

$$B^a(e|\lambda') \triangleq \sum_{t \in \mathcal{T}^a} A^a(e, t|\lambda') \quad (20)$$

We give below the solution for each parameter:

- the *a priori* probability of environment and *a priori* probability of trajectory:

$$\eta_e = \frac{\sum_{a \in \mathbb{P}} B^a(e|\lambda')}{\sum_{a \in \mathbb{P}} |\mathcal{X}^a| |\mathcal{Y}^a|}, \quad \zeta_{e,t}^a = \frac{A^a(e, t|\lambda')}{B^a(e|\lambda')} \quad (21)$$

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

- the mean vector and the covariance matrix of the environment observation distribution:

$$m_e^{(y)} = \frac{\sum_{y \in \mathcal{Y}} C_y^e y}{C_y^e} \quad (22)$$

$$\Sigma_e^{(y)} = \frac{\sum_{y \in \mathcal{Y}} C_y^e (y - m_e^{(y)}) (y - m_e^{(y)})^\#}{\sum_{y \in \mathcal{Y}} C_y^e} \quad (23)$$

where $\#$ stands for transposition operation and:

$$C_y^e \triangleq \frac{p(y|e, \lambda')}{p(y|\lambda')} \quad (24)$$

- the mean vector and the covariance matrix of the acoustic observation distribution:

$$m_{i,e,t}^{(x)a} = \frac{\sum_{X \in \mathcal{X}^a} D_{i,e,t}^a x_i}{\sum_{X \in \mathcal{X}^a} D_{i,e,t}^a} \quad (25)$$

$$\Sigma_{i,e,t}^{(x)a} = \frac{\sum_{X \in \mathcal{X}^a} D_{i,e,t}^a (x_i - m_{i,e,t}^{(x)a}) (x_i - m_{i,e,t}^{(x)a})^\#}{\sum_{X \in \mathcal{X}^a} D_{i,e,t}^a} \quad (26)$$

where:

$$D_{i,e,t}^a \triangleq \frac{p(X|t, e, \lambda')}{p(X|\lambda')} \quad (27)$$

4. EXPERIMENTS AND RESULTS

4.1. Task description

Experiments deal with a 1011 word vocabulary noisy continuous speech recognition task. 140 short sentences read by 10 French speakers (2 females, denoted brs and syc in the following) have been used for training. For testing, 160 sentences (60 common to all speakers and 100 specific to each speaker) were recorded. In average, there are about 70 observations per phoneme for each speaker in training. The recording system used a SUN desktop omnidirectional microphone, mounted on the terminal. The distance between a speaker and the microphone was about 40 cm in average and 60 cm for some speakers (jel, crm), resulting in an SNR of about 15 dB. Speech is sampled at 16 kHz. 13th order mel-cepstral vectors were computed every 10 ms with an analysis window of 32 ms. The task is difficult because of insufficient training data and noisy recording conditions, and because between-word pauses are not modeled by our grammar.

The acoustic observation X is a sequence of mel-cepstral vectors, y is supposed to capture speaker information and is computed as the average of the mel-cepstral vectors over the whole utterance. The language model used has a word-pair equivalent perplexity of 29. The acoustic models are initialized with the phoneme segments produced by the method of [5] and re-trained by one iteration of segmentation-reestimation. The number of mixtures ($|\mathcal{T}^a|$) is proportional to the number of observations in the training data. In MSTM as well as in ED-MSTM, the number of states Q is fixed to 5. For the experiments, 32 context-independent phone models, including one silence model, are build. Up to 4 ($|\mathcal{E}| = \Delta$) environment components are used for ED-MSTM, initialised by LBG algorithm [8]. Duration probability estimation and sentence recognition use the same technique as in the VINICS recognition system [6].

4.2. Experiment design

We tested the system in three configurations:

- speaker-dependent mode (SD): we set $|\mathcal{E}| = 1$ and trained a system for each speaker,
- multi-speaker mode (MS): we set $|\mathcal{E}| = 1$ and trained an unique system with the data from all speakers,
- speaker-clustering mode (SC): we set $|\mathcal{E}| = 4$ and trained an unique system with the data from all speakers.

In (SD) each system is tested speaker-dependently, and in (MS) and (SC) the system is tested on the whole speaker population.

Results in terms of word recognition accuracy are given in Table 1. The total number of pdfs used for each configuration are similar. Most errors are due to inter-word pauses which are not covered by our language model. The (MS) system gives 88.0% word accuracy (with 58 Del., 550 Sub. and 162 Ins. over 6418 words), with a 95% confidence interval of 87.2% – 88.8%. We observe that (SC) gives the highest recognition rate (89.64% word accuracy with 105

Del., 477 Sub. and 83 Ins. over 6418 words, with a 95% confidence interval of 88.9% – 90.4%), as expected, and that the use of environmental modeling reduced by about 15% the error rate produced by the multi-speaker system. This represents a statistically significant improvement.

5. CONCLUSION

To cope with long term variabilities which do not change within one utterance, we propose a new model of hidden environment sources. Each environment source models the short term variability by means of a mixture stochastic trajectory model. The estimation of the involved parameters is derived under the EM framework. Experimental results show that a recognizer based on this new model gives a higher recognition accuracy, compared to a system trained in a multi-speaker mode with similar number of parameters.

An idea related to the environment-dependent acoustic modeling proposed in this paper is the gender-dependent acoustic modeling [4], where two sets of models (male/female) are trained and the recognizer searches the best solution between the two sets of models. Our formulation allows any number of model sets ($|\mathcal{E}|$) with an ML estimate of model parameters using EM, and can also be applied to other types of environments such as noisy recording condition.

Compared to MSTM, in environment-dependent MSTM, the trajectory models are optimally grouped (in ML sense) into clusters according to an environment measurement. This technique prevents trajectories of different groups of speakers from being mixed up, thus improving the discriminability of the acoustic modeling.

The idea of environment modeling is applicable to other modeling techniques such as Hidden Markov Models.

Suitable parameter tying in the model can help to find a better trade-off between modeling accuracy and reliability. Several tying configurations can be envisaged. For instance, tying parameters at trajectory level, i.e.: all trajectories are pooled into a base shared by each of the phoneme symbols, or at Gaussian pdf level across symbols, i.e.: the pdf of the Q -points are assumed independent of symbol and trajectory. These directions are now being investigated.

6. REFERENCES

1. M. Afify, Y. Gong, and J. P. Haton. Stochastic trajectory model for speech recognition: an extension to modelling time correlation. *In Proc. of European Conference on Speech Communication and Technology*, 1:515–518, September 1995. Madrid, Spain.
2. C. Cerisara, Y. Gong, and J. P. Haton. Reconnaissance de la parole continue par le modèle STM polynômial. *In Actes des 21-èmes 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.*

speaker	SD	MS	SC
brs	92.30	87.96	92.13
crm	86.40	85.65	83.84
jel	87.58	72.15	79.03
lar	94.89	96.06	95.04
ols	96.90	95.60	96.41
sat	93.94	80.69	81.67
std	91.08	90.77	89.69
syc	94.08	85.60	90.24
vil	93.27	92.41	92.55
yig	95.57	90.99	94.50
AVG	92.60	88.00	89.64

Table 1: Word accuracy rates as function of speakers and test modes (SD - speaker-dependent mode, MS - multi-speaker mode, SC - speaker-clustering mode)

- on Acoustics, Speech and Signal Processing, 1:557–560, April 1994. Adelaide, Australia.
5. Y. Gong and J. P. Haton. Iterative transformation and alignment for speech labeling. *In Proc. of European Conf. on Speech Communication and Technology*, 3:1759–1762, September 1993. Berlin, Germany.
6. 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.
7. I. Illina and Y. Gong. Stochastic trajectory model with state-mixture for continuous speech recognition. *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.