A Study on Continuous Chinese Speech Recognition Based on Stochastic Trajectory Models

Xiaohui MA Yifan GONG** Yuqing FU* Jiren LU* Jean-Paul HATON***

* Department of Radio Engineering, Southeast University, Nanjing, 210096, P.R.China
** CRIN/CNRS - INRIA-Lorraine, BP 239, 54506 Vandoeuvre, France

ABSTRACT

This paper first introduces the theory of Stochastic Trajectory Models (STMs). STM represents the acoustic observations of a speech unit as clusters of trajectories in a parameter space. The trajectories are modeled by mixture of probability density functions of random sequence of states. Each state is associated with a multi-variate Gaussian density function, optimized at state sequence level. The effect of not using the HMM assumptions in STM is that STM can exploite information, such as time correlation within an observation sequence, which is hidden by HMM assumptions.

After analyzing the characteristics of Chinese speech, the acoustic units for recognizing continuous Chinese speech taking advantage of Stochastic Trajectory Models are discussed and phone-like units, which are similar to or smaller than Initial-Final-like units, are suggested. The total number of the phone-like units (about 50) is the smallest in almost all Chinese speech recognition systems. Consequently, the training database can be very small.

The performance of continuous Chinese speech recognition based on STM is studied using the VINICS system. The experimental results demonstrate the efficiency of STM and the consistency of phone-like units.

1. INTRODUCTION

Hidden Markov modeling (HMM) is widely used in continuous speech recognition, which models speaking rate by state transition probabilities and acoustic variability by state-dependent observation probability densities. To make computation efficient, HMM algorithms assumes:

Assumption 1 Transition probability is independent of observation sequences and is time-invariant;

Assumption 2 State observation pdf is independent of past state;

Assumption 3 State observation pdf is independent of the previous observations.

As speech is produced by a continuously-changing system, these assumptions are obviously unrealistic. In a parametric space (i.e. cepstral space) speech signal can be represented as a point, which moves as articulatory configuration changes. The sequence of moving points is called the trajectory of speech [1]. Due to the three unrealistic assumptions, present HMM cannot preserve trajectory information. Particularly, HMM results in trajectory folding phenomenon which leads to a poor discriminability in complex contexts [1]. To compensate this, several methods have been tried, such as dynamic coefficients [2], bigram constraints between observation vectors [3] and context-dependent acoustic sub-word models [4].

In this paper, speech trajectories are modeled by stochastic trajectory models (STMs) [1], which represent the acoustic observations of a speech unit as clusters of trajecotries in a parameter space. The trajectories are modeled by mixture of probability density functions of random sequence of states. Each state is associated with a multi-variate Gaussioan density

function, optimized at state sequence level. Conditional trajectory duration probability is integrated in the modeling. After analyzing the characteristics of Chinese speech, the acoustic units for recognizing continuous Chinese speech taking advantage of Stochastic Trajectory Models are discussed and phone-like units, which are similar to or smaller than Initial-Final-like units, are suggested. The total number of the phone-like units (about 50) is the smallest in almost all Chinese speech recognition systems. The performance of continuous Chinese speech recognition based on STM is studied using VINICS system. The experimental results demonstrate the efficiency of STM and the consistency of phone-like units.

2. STOCHASTIC TRAJECTORY MODELS[1]

Let X_n be a sequence of Q vectors (points in parameter space) centered at time slot n:

$$X_n = x_{n-\frac{Q}{2}}, x_{n-\frac{Q}{2}+1}, \ldots, x_n, \ldots, x_{n+\frac{Q}{2}-1}$$

Each speech unit symbol s is associated with a stochastic model T_s of trajectories. X_n is assumed to be obtained by re-sampling a sequence of d frames:

$$[0,d) \xrightarrow{f(i,d,Q)} [0,Q)$$

then

$$p(X_n,d,s) = p(X_n|d,s)P(d|s)P(s)$$

Let T_s be represented as a mixture of component trajectories t_k:

$$p(X_n|d,s) = \sum_{k \in T_s} p(X_n|t_k,d,s)P(t_k|d,s)$$

For computational efficiency, we make the assumption that each of the Q points of the component trajectory t_k is produced by an independent distribution, the pdf of X_n given t_k, d and s is modelled as a multivarite Gaussian distribution, with mean $m_{k,i}^s$, covariance matrix $\Sigma_{k,i}^s$ and state weight ω_i^s:

$$p(X_n|t_k,d,s) = \prod_{i=0}^{Q-1} p(x_{n-f(i,d,Q)}|t_k,d,s)$$
$$= \prod_{i=0}^{Q-1} N(x_{n-f(i,d,Q)}; m_{k,i}^s, \Sigma_{k,i}^s)^{\omega_i^s}$$

The probability of speech unit s given X_n and d is therefore:

$$P(s|X_n,d) = \frac{p(X_n,d,s)}{p(X_n,d)} = \frac{p(X_n,d,s)}{\sum_s p(X_n,d,s)}$$

Let a particular sentence w be made up of $L(w)$ symbols:

$$w = s_1, \ldots, s_h, \ldots, s_{L(w)} \qquad h \in [1, L(w)]$$

Let speech segment $[n_{h-1}+1, n_h]$ correspond to symbol s_h and $n_0 = 0$.

Then, we optimize the probability of sentence w

$$\Theta(w) = \max_{n_1, \ldots, n_{L(w)}} p(n_1, \ldots, n_{L(w)})$$
$$= \max_{n_1, \ldots, n_{L(w)}} \prod_{h=1}^{L(w)} P(s_h | X_{\frac{n_{h-1}+n_h+1}{2}}, n_h - n_{h-1})$$

Sentence recognition consists in evaluating the maximized cumulated probability for all possible sentences, and assign the most probable sentence as the recognized sentence:

$$\arg\max_w \Theta(w)$$

The fundamental difference between STM and mixture HMM approach is: in STM, the mixture of densities is defined on the state sequence whereas in HMM it is defined on individual states. The effect of not using the HMM assumptions in STM is that STM can exploite information, such as time correlation within an observation sequence, which is hidden by HMM assumptions. By explicitly modeling a speech trajectory as a mixture of sequences of states, STM can avoid the trajectory folding phenomenon. And furthermore, STM uses an accurate explicit phoneme duration probability modeling and allows integration of state weight ω_i^s on the contribution of each observation vector within a segment to the probability of the segment. These duration modeling and

weighting, which can improve considerably recognition accuracy, can not be easily achieved within HMM framework. Therefore, STM provides a more in-depth modeling of continuous speech signals.

3. ACOUSTIC UNITS OF CHINESE SPEECH RECOGNITION

Good acoustic units for speech recognition should be consistent. Spoken Chinese is a syllabic language and each syllable corresponds to one character. The total number of phonologically allowed syllables (regardless of tones) of Chinese is about only 410. One natural and direct way to model Spoken Chinese consists therefore in selecting syllable units as acoustic units for recognition. But this selection will result in some problems of inconsistency since Chinese syllables consist of 37 very confusing sets which are difficult to recognize. To overcome these problems, almost all state-of-the-art Chinese speech recognition systems use Initial-Final-like units as acoustic units owing to the fact that each Chinese syllable can be decomposed into an Initial-Final format. However, present HMM cannot preserve trajectory information and the Initial parts are in general shorter and unstable. Thus, the HMM recognition systems have to use syllable-dependent units[5], particularly syllable-dependent Initial units, which leads to increasing the total number of the acoustic units.

Now that STM can provide a more in-depth modeling of speech signals, we can use Initial-Final-like units as acoustic units directly. Moreover, each Chinese syallable can be decomposed as follows:

$$\text{Syllable} = \underbrace{(\text{Consonant})}_{\text{Initial-part}} + \underbrace{(\text{Vowel})\ \text{Vowel}\ (\text{Vowel, Nasal})}_{\text{Final-part}}$$

so the complex-Final-part can be decomposed into either a vowel followed by another vowel or a vowel followed by a nasal. Such as complex-Final-part 'ua' can be decomposed into vowel 'u' and vowel 'a', 'ing' can be decomposed into vowel 'i' and nasal 'ng'. Then, these phones will be merged in a specially designated way. For example, the vowel 'u' in 'ua' and 'u' in 'uai' will be merged into one vowel entitled 'u_a'. We call the acoustic units formed by this approach phone-like units, which are similar to or smaller than Initial-Final-like units. The total number of the phone-like units (about 50) is the smallest in almost all Chinese speech recognition systems. Consequently, the training database can be very small.

4. EXPERIMENTAL RESULTS

The performance of continuous Chinese speech recognition based on STM was studied using the VINICS system[6].

The validation experiments are based on task-independent acoustic training, i.e. the vocabulary of the training text has been designed to have little coverage over that of recognition text. The text of training utterances consists of 120 sentences randomly selected from a Chinese textbook. The text of testing utterances consists of 50 sentences with about 180 words for a Chinese software command-and-control application. Speech is sampled at 11.025kHz, blocked each 10ms with 256 points window, and parametrized using 13 mel-cepstral coefficients.

50 context-independent phone-like-units models are used for the experiment. Acoustic models are trained using boot-strapping technique. A language model based on context-free grammars (CFG) is used. To find the best sentence, the recognition system performs beam search with N-best sentences as final result. In counting errors, only the top one sentence is used.

The speaker-dependent experimental result is shown in Table-1. Sub, Del, Ins, Acc are respectively the number of substitutions, deletions, insertions and accuracy.

Speaker	%Sub	%Del	%Ins	%Acc
1	1.1	0	1.1	97.8
2	2.8	0	0.5	96.7

Table 1: word recognition result

5. CONCLUSION

In this paper, the theory of Stochastic Trajectory Models (STMs)

is introduced. STM represents the acoustic observations of a speech unit as clusters of trajectories in a parameter space. The trajectories are modeled by mixture of probability density functions of random sequence of states. Each state is associated with a multi-variate Gaussian density function, optimized at state sequence level. The effect of not using the HMM assumptions in STM is that STM can exploite information, such as time correlation within an observation sequence, which is hidden by HMM assumptions.

After analyzing the characteristics of Chinese speech, the acoustic units for recognizing continuous Chinese speech taking advantage of Stochastic Trajectory Models are discussed and phone-like units, which are similar to or smaller than Initial-Final-like units, are suggested. The total number of the phone-like units (about 50) is the smallest in almost all Chinese speech recognition systems. Consequently, the training database can be very small.

The experimental results demonstrate the efficiency of STM and the consistency of phone-like units.

6. REFERENCES

1. Gong, Y.-F., and Haton, J.-P., "Stochastic Trajectory Modeling for Speech Recognition", *Proc.ICASSP'94*, pp.57-60, 1994.
2. Furui, S., "Speaker-Independent Isolated-Word Recognition Using Dynamic Features of the Speech Spectrum", *IEEE Trans.ASSP*, Vol.34, No.1, pp.52-59, 1986.
3. Paliwal, K.K., "Use of Temporal Correlation Between Successive Frames in a Hidden Markov Model Based Speech Recognizer", *Proc.ICASSP'93*, pp.215-218, 1993.
4. Lee, K.-F. "Large Vocabulary Speaker-Independent Continuous Speech Recognition : The SPHINX System", PhD thesis , Carnegie-Mellon University, 1988.
5. Hon, H.-W., et al, "Towards Large Vocabulary Mandarin Chinese speech Recognition", *Proc.ICASSP'94*, pp.545-548, 1994.
6. Gong, Y.-F., and Haton, J.-P., "VINICS: A Continuous Speech Recognizer Based on a New Robust Formulation", *Proc.ECSCT*, pp.1221-1224, 1991.