ACOUSTIC OBSERVATION CONTEXT MODELING IN SEGMENT BASED SPEECH RECOGNITION

Máté Szarvas

Technical University of Budapest
szarvas@bme-tel.ttt.bme.hu

Shoichi Matsunaga

NTT Human Interface Labs.
mat@nttspch.hil.ntt.co.jp

ABSTRACT

This paper describes a novel method that models the correlation between acoustic observations in contiguous speech segments. The basic idea behind the method is that acoustic observations are conditioned not only on the phonetic context but also on the preceding acoustic segment observation. The correlation between consecutive acoustic observations is modeled by polynomial mean trajectory segment models. This method is an extension of conventional segment modeling approaches in that it not only describes the correlation of acoustic observations inside segments but also between contiguous segments. It is also a generalization of phonetic context (e.g.,triphone) modeling approaches because it can model acoustic context and phonetic context at the same time. In a speaker-independent phoneme classification test, using the proposed method resulted in a 7–9% reduction in error rate as compared to the traditional triphone segmental model system and a 31% reduction as compared to a similar triphone HMM (hidden Markov model) system.

1. INTRODUCTION

1.1. Intrasegment Correlation

Most current speech recognition systems are based on frame-based measurements. HMM systems make the further assumption that these frames are statistically independent, given the state. The reason for making this assumption, however, has to do with computational efficiency in implementing practical systems, not with theoretical or experimental evidence. Actually, we know experimentally that the frames of a speech segment corresponding to one phonetic event are highly correlated. Consequently, the price of computational efficiency is degraded recognition accuracy.

1.2. Segmental Models

To alleviate this problem, several systems have been proposed recently that use segment-based measurements (usually calculated from frame-based ones) to jointly model the observations corresponding to one phonetic segment. Such models are usually called *segment models*, of which a comprehensive overview is presented in [5].

1.3. Intersegment Correlation

Segment models address the problem of intrasegment correlation. Perceptual experiments, however, indicate that significant correlation exists not only inside phonemes but also between neighboring ones [2].

The most important clue for the identity of certain consonants is the spectral change (formant transition) in the preceding and following vowels.

Context dependency is often modeled by using separate models for the same phoneme as a function of the adjoining phonemes (triphone models) [4]. In this paper, we call this method *phonetic context* modeling.

An analysis of the basic maximum a-posteriori formulas of automatic speech recognition suggests that it is possible to model not only phonetic context, but also *observation context*. That is, to determine the probability of a segment observation, we can use both the identity of the adjacent phonemes and the acoustic observations corresponding to them.

In the next section we derive the basic equations for modeling observation context. The formalism will be introduced using class-conditional probabilities. After deriving the basic formulas, we shall describe one way of implementing them using polynomial mean-trajectory segment models. In Section 4 we review the results of a phoneme classification experiment with observation context modeling and we summarize our paper in Section 5.

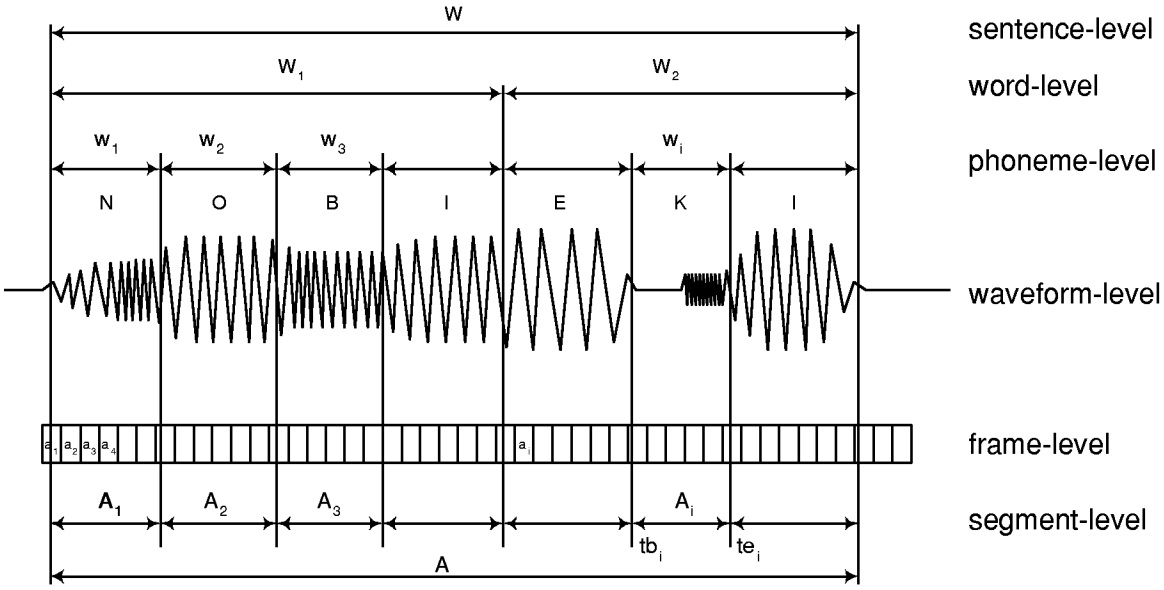
2. MATHEMATICAL FORMULATION

The most common approach to continuous speech recognition is to find the word (phoneme) sequence, W, which maximizes the joint probability (likelihood) of the acoustic observation, A, and the word sequence, W.

$$W^* = \arg\max_W p(A, W) \quad (1)$$

$$= \arg\max_W p(A|W)p(W). \quad (2)$$

In the usual triphone-based approach $P(A|W)$ is factorized



tb_i, te_i	begin and end times for the ith segment observation
a_t	acoustic observation at time instant t (frame)
A_i	the segment observation corresponding to the ith subword unit, w_i, $A_i = \{a_{tb_i}, a_{tb_i+1}, \ldots, a_{te_i}\}$
A	segment observations for a whole utterance, $A = A_1^N = \{A_1, \ldots, A_N\}$
w_i	the ith subword unit (e.g., phoneme) in the utterance (sentence)
W_k	kth word in the utterance
W	subword-unit (or word) sequence for the whole utterance, $W = w_1^N = W_1^K$

Figure 1. Summary of the notation used in the paper

to

$$P(A|W) = \prod_i P(A_i|w_{i-1}^{i+1}), \tag{3}$$

where A_i is the i-th acoustic segment observation and the conditioning term is the triphone centered around phoneme w_i in W.

Because our method does not assume the independence of the contiguous segment observations A_i and A_{i-1}, $P(A_i)$ is conditioned not only on the triphone but also on the preceding segment observation:

$$P(A|W) = \prod_i P(A_i|A_{i-1}, w_{i-1}^{i+1}). \tag{4}$$

3. IMPLEMENTATION

Class-conditional probabilities are usually estimated by density estimation methods, which can not be directly conditioned on continuous parameters. Therefore, we have to use the basic definition of conditional probabilities in order to model the dependence on A_{i-1} in Eq. (4):

$$P(A_i|A_{i-1}, w_{i-1}^{i+1}) = \frac{P(A_{i-1}, A_i|w_{i-1}^{i+1})}{P(A_{i-1}|w_{i-1}^{i+1})}. \tag{5}$$

The terms in the numerator and denominator are both usual triphone probabilities and standard methods exist for estimating them. However, it is important not to use a distribution in the numerator that assumes the independence of A_i and A_{i-1}, because our purpose is to model the correlation between them.

Our implementation uses polynomial segment (mixture) models to estimate this equation. Separate models are used to estimate the numerator and denominator. We have a set of segment observations $\{A_i\}$, in which each element A_i is the realization of the same triphone. To estimate the numerator of Eq. (5), we make the new set $\{A_i^*\} = \{(A_{i-1}, A_i)\}$, in which each element, A_i^*, is the concatenation of A_i from the original set and the acoustic segment observation, A_{i-1}, preceding it in the original utterance. Using this set of modified segment observations and the estimation algorithm described in [1, 3], we can obtain a model that implements the numerator of Eq. (5). Because a segment model obtained in this way does not assume the independence of A_i and A_{i-1}, it can model the correlation between them.

The denominator is estimated in a similar way, here the modified set, $\{A_i^*\} = \{A_{i-1}\}$, consists of the segment observations preceding each A_i in the original set.

$P(A_i|A_{i-1}, w_{i-1}^{i+1})$ is then obtained by taking the quotient of the likelihoods produced by the two models.

The computational complexity of the method is about three times that of the basic segment modeling approach without acoustic context modeling. Calculating the numerator requires about two times the computation because the length of (A_{i-1}, A_i) is twice that of the length of A_i, on average. Calculating the denominator requires the same amount of computation as the original method because the length of A_{i-1} is the same, on average, as the length of A_i. This value can, however, be decreased by using only the last few frames of A_{i-1}, both in the numerator and in the denominator. Using only the frames close to the segment boundary is warranted by the assumption that the correlation is the most significant between the frames close to the transition region.

4. EXPERIMENTAL EVALUATION

In order to evaluate the practicality of the proposed method, speaker-independent phoneme classification experiments were performed using the "ATR 520 Important Japanese Words" database. Fifteen male and fifteen female speakers were used for training and another five of each gender were used for testing.

The speech was originally sampled at 12 kHz. Every 10 milliseconds a vector of 13 Mel-warped cepstral coefficients was computed using a 25-millisecond window of the speech.

In some of the experiments, in addition to these "static" coefficients, the so-called delta and acceleration coefficients were also used. These coefficients were calculated using the regression method with ± 2 frames of data.

After a word was parameterized, the mean vector was determined and subtracted from the parameter vector of each frame (cepstral mean removal) in order to increase the robustness against speaker and channel variations.

The triphone models used in some of the experiments were of the generalized triphone type described in [4].

The results of the experiments are summarized in Tables 1–4.

The models had three mixtures when not indicated otherwise. The HMM models had three states with a diagonal covariance matrix. The polynomial segment models (PSMs) had second order of mean trajectory polynoms when not indicated otherwise, while the variance trajectory polynom order is displayed in the tables explicitly. We note that in Table 1 the PSM models had a smaller number of free parameters than the HMMs, because of the constant variance trajectory. The following duration models were evaluated: no explicit duration model (NO), normal duration distribution (DN), and the gamma duration distribution (DG). The HMMs always used the inherent exponential duration model.

The segment models with observation context modeling took into account only the last 20 ms of the preceding segment. This value was chosen so that the entire transition region could be included and the distant acoustic data avoided.

Two ways of using acoustic context were evaluated. The first way implemented only the numerator of Eq. (5). That is it calculated

$$P(A_{i-1}, A_i|w_{i-1}^{i+1}). \quad (6)$$

The second way implemented the entire Eq. (5). The two methods are indicated in the tables by their equation numbers.

Table 1 compares the performances of the HMM and the different PSM models. The following conclusions can be drawn from the results. The use of the segment model decreases the error rate, as compared to that of the HMM, even in the case of a smaller number of free parameters (the PSMs used a static variance polynom). Using Eq. (6) to model acoustic context decreases the error rate significantly in the case of triphone (CD) models but hardly at all in the case of monophone (CI) models (5.71% vs. 0.52%). However, although Eq. (5) is suitable for continuous word recognition, using it to model acoustic context always increases the error rate in both cases. Using an explicit duration model decreases the error rate of the PSMs further, and a normal distribution seems to be better for this purpose than the gamma distribution.

Table 2 displays a more detailed comparison of the triphone HMM and different triphone PSMs. All of the PSMs evaluated in this table produced a smaller error rate than the HMM. It is apparent that increasing the variance polynom order decreased the error rate. It is also confirmed that the normal duration distribution gives the largest improvement, although the gamma distribution is also better than no explicit model. Finally, when Eq. (6) was used to model acoustic context the error rates in all cases decreased, and the opposite was true when Eq. (5) was used.

The effect of the length of the acoustic observation contect is shown in Table 3. Here triphone PSMs are compared which use an acoustic context between 0 and 50 milliseconds. The optimum size of the acoustic context is about 30 milliseconds. This confirmes that only acoustic data close to the transition region is useful.

Finally, Table 4 compares the error rates of a set of models with the same number of free parameters for easier comparision. We can see that the use of the segment model resulted in a 25–26% lower error rate as compared to the HMM. Using of acoustic context decreased the error rate by another 7–9%.

5. SUMMARY

This paper proposed a method for modeling the correlation of acoustic parameters between adjacent segments. The results of the experiments indicate that for the given phoneme classification task, using an observation context consistently decreases the error rate. The error rate of the triphone polynomial segmental model system was 9–13%

Table 1. Speaker-independent phoneme classification error rates (%). Monophone (CI) and triphone (CD) models, as indicated. 13 dimensional feature vectors (mel-cepstrum). Constant variance trajectory polynoms (1 free parameter).

Model type	Dur. model	Obs. ctxt.	Parameters	
			CI	CD
HMM	NO	NO	31.98	15.47
PSM	NO	NO	28.76	12.78
		Eq. (6)	28.61	12.05
		Eq. (5)	29.88	13.22
	DG	NO	28.14	11.81
	DN	NO	27.42	11.69

lower when an observation context was used than when one was not. This improvement is significant and verifies the practicality of the new method for phoneme classification.

6. ACKNOWLEDGEMENTS

The authors wish to thank the useful advices of their collegaues during this research. We are especially grateful to Dr. Yasuhiro Minami and Dr. Satoshi Takahashi at the NTT Human Interface Laboratories and Prof. Péter Tatai at the Technical University of Budapest.

7. REFERENCES

[1] T. Fukada, Y. Sagisaka, and K. K. Paliwal. Model parameter estimation for mixture density polynomial segment models. In *ICASSP*, pages 1403–1406, 1997.

[2] Sadaoki Furui. On the role of spectral transition for speech perception. *J.Acoust.Soc.Am.*, 80(4):1016–1025, October 1986.

[3] H. Gish and K. Ng. A segmental speech model with applications to word spotting. In *ICASSP-93*, pages II/447–450, 1993.

[4] Kai-Fu Lee. *Automatic Speech Recognition: the developement of the SPHINX system.* Kluwer Academic Publishers, Norwell, Massachusetts 02061, 1989.

[5] M. Ostendorf, V. V. Digalakis, and O. A. Kimball. From HMMs to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*, SAP-4(5):360–378, September 1996.

Table 2. Speaker independent phoneme classification error rates (%). All the models were triphone models. 39 dimensional feature vectors (cepstrum, Δ and $\Delta\Delta$ cepstrum). The variance polynom order was between 0 and 2, as indicated.

Model type	Var. poly.	Obs. ctxt.	Duration Model		
			NO	DN	DG
HMM, 3 states		NO	13.57	—	—
PSM	constant	NO	10.93	10.48	10.52
		Eq. (6)	10.19	9.50	9.54
		Eq. (5)	11.61	10.93	11.05
PSM	linear	NO	10.89	10.27	10.40
		Eq. (6)	10.05	9.31	9.40
		Eq. (5)	11.61	10.94	11.04
PSM	quadratic	NO	10.93	10.18	10.28
		Eq. (6)	10.00	9.31	9.89
		Eq. (5)	11.58	10.76	10.84

Table 3. Effect of the length of acoustic context on speaker-independent phoneme classification error rates (%). Cepstrum, Δ and $\Delta\Delta$ cepstrum. Quadratic variance polynom. Acoustic context was taken into account using Eq. (6).

Model type	Observation context size	Duration Model	
		NO	DN
CD	0 ms	10.93	10.18
	10 ms	10.26	9.63
	20 ms	10.00	9.31
PSM	30 ms	9.84	9.16
	40 ms	10.09	9.60
	50 ms	10.70	10.09

Table 4. Speaker-independent phoneme classification error rates (%). All the models were triphone models. 39 dimensional feature vectors (cepstrum, Δ and $\Delta\Delta$ cepstrum). The variance polynom was quadratic. The PSMs used a normal duration distribution.

Model type	Obs. ctxt.	Parameters	
		static	static+Δ+$\Delta\Delta$
HMM	NO	15.47	13.57
PSM	NO	11.48	10.18
PSM	Eq. (6)	10.67	9.31