DATA REFINING HMM, A NEW APPROACH TO HMM BASED SPEECH RECOGNITION SYSTEM IMPROVEMENT

43.72.Ne

Razzazi,Farbod[1,2]; Sayyadian Abolghassem[1]
[1] Amirkabir University of Technology, Electrical Engineering Department. Information Processing Lab.
[2] Azad Islamic University, Research and Sciences Campus, Faculty of Engineering, Electrical Engineering Department.
[1] Hafez Ave., Abureyhan building, Room 315, Tehran, Iran.
[2] Hesarak Ave. Faculty of Engineering, Room 204, Tehran, Iran
Tel: (+9821) 780-4872
Fax: (+9821) 790-7266
Email: r7623915@cic.aut.ac.ir

ABSTRACT

The accuracy of HMM based speech recognition systems is limited due to some HMM presumptions. In this paper, a decorrelation method named “data refining” has been applied to HMM input sequence. The main idea in this study is to produce a nearly independent vector sequence containing all of the speech information in addition to consistency with model assumptions. In this paper, A few data refining methods were proposed and examined and the results were compared with each other and with standard HMM. Data refining method shows 3% improvement on recognition rate and 30% improvement on recognition computational cost, much better than other SSM methods.

1. INTRODUCTION

HMM is one of the most successful models in speech recognition systems. This success is indebted to model’s analytical strength and also empirical tests [1-3]. Besides, there are some disadvantages in HMM systems that limit the performance of the HMM recognition system.

To overcome this limitation, it is appropriate to summarize the HMM motivation briefly. The first statistical modeling that may come to a mind is estimating a pdf to statistically describe the feature vectors of an utterance. But using a non-conditional pdf for feature vectors is based on the stationarity assumption, which is not apparently applicable to speech. HMM solves this problem by estimating a pdf for each partially stationary segment of a lingual unit. In other words, the nonstationary sequence of vectors is modeled with a sequence of stationary segments. A Markov modeling is applied to the problem of transferring from a stationary segment to other segments.

There are two main presumptions in HMM modeling, First it is assumed that feature vectors are independent and as a result, the probability of the vector sequence can be calculated as the multiplication of each vector.

Second, It is assumed that stationary segments are connected through a Markov model, which implicitly imposes a geometric distribution on each state’s duration that is not empirically true.

This paper investigates the feature extraction process from uniformly located frames. Elimination of this assumption may resolve both first and second assumptions, because considering the independence presumption in addition to Markov assumption is a source of modeling deficiency only when the features are not representing different acoustic events.

2. DATA REFINING HMM

HMM is considered as an efficient approach for short time stationary stochastic sequence modeling, but the independence assumption is directly shown off in calculating the utterance probability as the multiplication of vectors probabilities. Data refining is motivated from the idea that speech is a sequence of different acoustic random length events and each event can be represented with a single feature vector. The length of each event does not convey much information about the speech meaning, but it rather represents the prosody and speaker information. In standard HMM, this temporal variation has directly been represented in the likelihood function that is not appropriate. Data refining idea is similar to speech segmental coding idea that is used in speech compression.

This approach suggests that a segmental feature vector for each segment is enough to describe the segments information and increasing the vectors even degrades the HMM modeling and recognition rate. The sequence of these new segmental features can be used in calculating HMM likelihood function of the observation sequence.

$$P(O_{refined} \mid \lambda) = P\ P(O_{refined}(i) \mid \lambda, s_{ji})$$

There are several ways to select the refined vector. In this paper, three methods were tested and the results are compared with each other and also standard HMM. In refining method I, it is assumed that each segment is represented with an intra-segmental distance. The segment boundary is the first vector that it's distance is greater than the first vector of the segment. The first vector of the segment is chosen as the refined vector of the segment.

$$\text{If } d(X_i, X_{ref}) / \text{norm}(X_{ref}) > \text{threshold then } X_{ref} = X_i;$$

The sequence of vectors $\{X_{ref}\}$ is determined as the refined output sequence.

Refining method II is based on the idea that the mean of the vectors is a better statistical measure of the segment rather than the first vector:

$$\text{If } d(X_i, \text{Mean(segment)}) / \text{norm(Mean(segment))} > \text{threshold then } X_{ref} = \text{Mean(segment)};$$

The third refining method tries to combine these two methods. Segment boundaries are selected as the first method, but the refined vector is determined as the mean of segment's vector. In [12,13], an analogous idea is applied to explicit segmentation of speech to extract an acoustic lexicon, but it is not directly used in a recognition system. From other view, this method can be considered as a stochastic segment model parameter setting, because the inter-segmental dependencies were modeled with a unity probability given the refined vector.

$$P(O_{segment} \mid \text{segment}) = 1;$$

Data refining methods, in addition to model the intra-segmental dependencies reduce the Markov state transitions, which implicitly assumes a geometric probability density function for each state, because this assumption may limit the performance when there is an intra-dependent sequence of data. Indeed, data refining HMM, gets help from capability of automatic segmentation of HMM to cover the errors of explicit segmentation of data refining.

The key advantage of data refining HMM to general SSM is separation of intra-segmental dependencies and stochastic modeling of each segment. In this manner, the search space of parameters is limited to two separate axes instead of a plane. Hence, parameter estimation is much more achievable, more accurate and faster, considering the fact that the nature of speech has really separated these search spaces and this assumption will not thrown us away from the correct parameters.

The obtained results showed us that not only the recognition rate has been increased, but also the recognition time is reduced in comparison to classical HMM. Because the length of vector sequence is reduced and the refining time is ignorable.

There are some investigated approaches to overcome this problem. The first proposed solution has been adding delta and delta-delta coefficients to the vector [1-3]. In this manner, the intra-segmental correlation is partially modeled. All of the feature vectors used today are containing these coefficients.

A very general model to overcome these models is stochastic segmental modeling (SSM), but this approach has estimation problems due to very wide search space. Some higher order Markov models are also tested. [5-6]

The next try is releasing the geometric distribution assumption from state duration modeling that leads to segmental HMM models. [7-9] To limit the search space, there are some tying solutions that were utilizing SSM. [10-11]

In the next section, our approach to overcome HMM problems is discussed.

3. IMPLEMENTATION RESULTS

Data refining method has been tested on two databases, TIMIT and a comprehensive Farsi speech database. The algorithm has been used to classify the main vowels of both databases. The selection of vowels is due to their long duration and stability, so it is expected that data refining would be more effective on them.

The first used database is speaker independent labeled TIMIT database. Each speaker has uttered ten sentences including two identical sentences among all speakers. The TIMIT database and feature extraction characteristics are tabulated in table 1. Table 2 shows the statistical characteristics of the database.

Table 1. TIMIT Specifications

Parameter	Value
Number of Speakers (Training)	14 Women 24 Men
Number of Speakers (Test)	4 Women 7 Men
Sampling Frequency	16KHz
Quantization	12 Bit
Analysis Frame Length	6.25ms
Analysis Frame Shift	12.5ms
Pre-emphasis Coefficient	0.95
Features	MFCC+ LogEnergy+ ΔMFCC

Table 2. Number of English Phonemes

Vowel	aa	ae	Eh	Iy	Uh	Ux	ow
Train	254	246	316	519	47	99	183
Test	79	69	93	174	20	35	57

The second database is a speaker dependent labeled speech Farsi database which has gathered all possible Farsi syllable compositions (CV, CVC, CVCC) independent to language statistical characteristics. This database and the features extracted from that are characterized as table 3 and the number of phonemes is recorded in table 4 for this database. 80% of the data are used to train the system and the remaining data are considered as test set.

Table 3. Farsi Database Specification

Parameter	Value
Sampling Frequency	8 KHz
Quantization	16 Bit
Analysis Frame Length	6.25ms
Analysis Frame Shift	12.5ms
Pre-Emphasis Coefficient	0.95
Features	MFCC+ LogEnergy+ ΔMFCC

Table 4. Number of Farsi Phonemes

Vowel	a	@	E	I	o	u	W
Number	319	280	288	295	282	294	186

At the first stage, a standard CDHMM based recognition system was developed with the parameters presented in table 5 and at the next step, seven main vowels of both databases are trained to HMM system.

Table 5. HMM Specifications

Parameter	Value		Parameter	Value
Number of States	3		Initialization	K-means
Number of Mixtures	4		a-matrix Initialization	Constant
Training Method	Multiple Observation Baum-Welch		Guassian Mixtures	Diagonal
			Other Specifications	Scaling + Logarithmic Calculations

The standard CDHMM recognition results are presented in table 6, on both training and test sets. Average and weighted average recognition rates are also calculated. The weighting has been performed based on the frequency of each phoneme in the data set.

Table 6. Vowel Recognition Results on TIMIT and Farsi Database

Vowel	Database	aa	Ae	eh	Iy	Uh	Ux	ow	Average	Weighted
		(a)	(@)	(e)	(i)	(o)	(u)	(w)		Average
Train	TIMIT	91.34	90.46	49.68	90.37	76.6	69.4	78.14	78	80.21
Test	TIMIT	87.61	87.34	66.67	91.95	60	62.86	66.67	74.87	80.03
Train	Farsi	91.1	95.2	94.6	98.1	76.5	87.1	50.7	84.74	********
Test	Farsi	67.9	94.9	84.6	88.5	21.8	60.3	69.2	69.6	********

To test the proposed data refining methods, the first method was applied to HMMs trained by English vowels. It was empirically understood that data refining does not have good performance on training phase. It can be explained by the fact that variety reduction in the refining procedure may cause the HMM not to generalize the trained patterns.

So, data refining is just applied in recognition phase. Table 7 is the recognition rate obtained for different refining thresholds. Comparing the results with table 6, it is obviously shown off a 3%

improvement in recognition rate at optimum threshold. An improvement and then adecrease in recognition rate are observed. It may be explained by revealing the refining effect at the first and then missing the information when continue to increase the threshold. The recognition time is 30% reduced using the optimum threshold.

The recognition results using second and third refining method has been presented in tables 8 and 9. The results show that in spite of increasing the complexity of the refining algorithm, the performance has ever been decreased. The recognition time has been increased too, because of more computational cost of the methods. This phenomenon may happen because different data typed have been utilized in training and test procedure. The first method that seems to be more effective has been applied to Farsi trained HMM system and the results have been presented in Table 10.

Table 7. Recognition Results Using Refining Method I

Phoneme	Th	Aa	ae	eh	iy	Uh	ux	ow	Average	Weighted Average
Train +Test	0.25	92.52	87.28	49.68	93.64	80.85	67.16	79.78	78.70	80.82
Test	0.25	87.24	79.75	62.37	93.68	65.00	62.86	68.42	74.30	79.52
Train	0.25	92.52	86.71	69.30	93.06	70.22	64.93	77.60	79.19	83.32
Test	0.25	89.87	78.48	82.80	93.10	60.00	60.00	63.16	75.35	82.12
Train	0.4	90.16	89.88	71.52	94.03	72.34	58.95	75.96	78.98	83.71
Test	0.4	88.61	83.54	84.95	93.68	60.00	57.14	66.67	76.37	83.43
Train	0.6	58.27	85.26	81.33	83.82	72.34	83.58	91.36	79.41	80.49
Test	0.6	43.04	75.95	84.95	88.51	55.00	68.57	80.70	70.96	75.98

Table 8. Recognition Results of Test Set Using Method II

Th	Aa	Ae	Eh	iy	uh	ux	Ow	Average	Weigthed Average
0.1	87.34	69.62	72.04	93.68	65.00	62.86	63.18	73.39	79.15
0.2	84.81	78.48	80.65	93.68	60.00	60.00	70.18	75.40	81.94
0.3	75.95	81.01	74.19	96.55	55.00	60.00	73.68	79.41	80.19

Table 9. Recognition Results of Test Set Using Method III

Th	aa	Ae	eh	iy	Uh	ux	ow	Average	Weighted Average
0.1	87.34	69.62	70.98	93.68	65.00	62.86	63.16	73.23	78.95
0.2	84.81	75.95	80.65	93.68	60.00	60.00	68.42	74.79	81.28
0.4	69.62	82.28	44.08	94.83	55.00	57.14	42.11	63.58	70.95

Table 10. Recognition Results On Farsi Database Using Refining Method I

Th	a	@	e	i	o	u	w	Average
0.4	71.3	96.8	87.8	91.0	31.5	58.4	70.1	72.41

4. CONCLUSION

In this paper, a data refining method has been presented to improve phoneme recognition rate. The results show a 3% improvement in recognition rate and 30% reduction in recognition time. The main idea in this paper is separation of segmentation of speech from statistical probability density estimation. This helps the stochastic modeling to calculate a more robust and valid likelihood function nearly independent of temporal warping of the utterance.

The research is in progress now and is concentrated on better refining methods and HMM substituting modeling.

5. REFERENCES

[1] L.R. Rabiner, “A Tutorial On Hidden Markov Models and Selected Applications in Speech Recognition”, Proceedings of The IEEE, Vol. 77, No. 2, Feb. 1989, PP. 257-286.

[2] C. Becchetti, L.R. Ricotti, “Speech Recognition, Theory and C++ Representation”, John Wiley & Sons, 1999.

[3] X. Huang, A. Acero, H. Hon,”Spoken Language Processing, A Guide to Theory, Algorithm, and System Development**”**, *Prentice Hall, 2001.*

[4] M. Ostendorf, V. Digalakis, O. Kimball, “From HMM To Segment Models: A Unified View of Stochastic Modeling of Speech Recognition”, IEEE Transactions on Speech and Audio Processing, Vol. 4, No. 5, Sep. 1996, PP. 360-378.

[5] J. A. Preez, “Efficient High Order Hidden Markov Modeling”, PhD Thesis, University of Stellenbosch, South Africa, 1997.

[6] J. D. Preez, “Algorithm for High Order Hidden Markov Modeling”, Proceeding of ICSLP, 1996, PP. 101-106.

[7] M. Russell, “A Segmental HMM for Speech Pattern Modeling”, Proceedings of IEEE ICASSP, Vol. 2, Apr. 1993, PP. 499-502.

[8] W. J. Holmes, M. J. Russel, “Modeling Speech Variability with Segmental HMMs”, Proceedings of ICASSP, 1996, PP. 447-450.

[9] W. J. Holmes, M. J. Russel, “Experimental Evaluation of Segmental HMMs”, Proceedings of ICASSP, 1995, PP. 536-539.

[10] M. J. Russel, W. J. Holmes, “Linear Trajectory Segmental HMMs”, IEEE Signal Processing Letters, Vol. 4, No.3, Mar. 1997, PP. 72-74.

[11] J. Goldberger, D. Burshtein, H. Franco, “Segmental Modeling using Continuous Mixture of Nonparametric Models”, IEEE Transactions on Speech and Audio Processing, Vol. 7, No. 3, May 1999, PP. 262-271.

[12] C. H. Lee, F. K. Soong, B.H. Juang, “A Segment Model Based Approach to Speech Recognition”, IEEE ICASSP, 1988, PP 501-504.

[13] S.A. Euler, B.H. Juang, C.H. LEE, F.K. Soong, “Statistical Segmentation and Word Modeling Techniques in Isolated Word Recognition”, Proceedings of IEEE ICASSP, Vol. 2, Apr. 1990, PP. 745-748.