An Evaluation of Language Identification Methods Based on HMMs

Seiichi Nakagawa* and Allan A. Reyes†

ABSTRACT

This paper describes two methods of language identification, both of which are based on HMMs(Hidden Markov Models). Here, we focused on the identification of 10 languages from the OGI Telephone Speech Corpus.

In the first method,a fully-structured(ergodic) HMM was trained for each language using text-independent speech samples from many native speakers. The likelihood for each language is calculated for the input speech using this HMM. In the second method, a universal ergodic HMM is trained using all the language data and with it, the most likely state sequence is computed for each language. The state sequence derived is processed and is used in the construction of trigram models for each language. The trigram model was used for modeling the phonotactics for each language.

Evaluation on the development test set of the OGI Corpus showed that combining these two methods gave a best performance of 58.5%.

1 INTRODUCTION

Automatic language identification by speech, while being similar to speaker identification, is one of the most difficult yet exciting areas in speech recognition because of individual differences in pronunciation and intonation and differences in the contents of the speech and thus, we should produce a good and robust language model for the variations caused by speakers or texts. Such language models can be of help in the realization of automatic dialing and translation systems and also in the construction of language training CAIs.

Even before the advent of the OGI Multi-Language Telephone Speech Corpus[1], there have been a number of researches in this field and consequently, several methods have been proposed to implement this language identification task. P.Henrich[2] studied the identification of words in three languages(German, English, French) with rules. He obtained almost the same performance as with the result using a neural network. A.House and E.Neuburg [3] used eight phonetic texts which were reduced to 4-character alphabets from

*Seiichi Nakagawa (中川 聖一): Professor, Department of Information Engineering, Toyohashi University of Technology

†Allan A. Reyes: Graduate Student, Department of Information Engineering, Toyohashi University of Technology

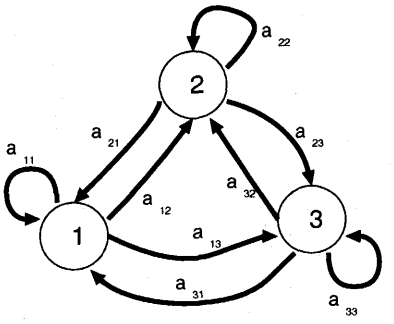

Figure 1: An Ergodic HMM with 3 states

26 alphabetical letters and these samples were reduced to form N-state statistical models of each language. However, they did not experiment on the identification.

By using acoustical signal sequence, not alphabetical sequence, R. Cole et al. [4] performed experiments using a neural network with acoustical distribution of stop consonants. Y.K. Muthusamy et al. extended this approach[5]. M.Sugiyama performed language identification experiments based on the distortion measure of the vector quantization[6]. M. Savic et. al proposed a language identification system based on pitch distribution and a linear predictive ergodic HMM[7]. Also,M.A.Zissman[8] performed language identification experiments based on the ergodic HMM and reported the same results as ours[11]. Recently, he reported in [9] a performance of 79.2% on the OGI Telephone Speech Corpus closed-set experiment by using a method based on language dependent phoneme recognition. T.J.Hazen et al[10] used the phonotactic, prosodic and acoustic properties of each language and reported a performance of 58.5% using the OGI database development test set. Their result is the best performance for the OGI database. Lamel et al[12] reported a performance of 59.7% on 10 second signals from the same test set using phone-based systems that generate a most likely sequence of phones for every reference language. We performed two-language identification experiments with this method but results were worse compared to the combination described here[13].

Even in a text-independent mode, the task of language identification is different from the task of speaker identification in that the former has to eliminate/absorb the acoustic variation caused by speaker individualities. We discuss some methods to tackle this.

2 MODELS AND METHODS

2.1 Ergodic HMM-based Identification Method [11, 14]

Each language has its own proper phonotactics and from this, the text-independent model should be a model with memory. HMM is one of the most suitable information models because it includes a multiple-order Markov model as a special case. In speech recognition, a left-to-right HMM has been used, although an ergodic HMM, as shown in Figure 1, should

be used in a text-independent language identification. Roughly speaking, each state of such HMM corresponds to a phonetic group such as front-vowels, back-vowels, nasals, voiced stops, unvoiced stops and so on. So, it is better to use feature vectors which are often used in speech recognition. We used here regression coefficients of spectral parameters over time as dynamic features.

A speaker independent and text-independent ergodic HMM is made for each reference language and the HMM parameters are estimated using the Baum-Welch algorithm. The likelihood of each speech input is accumulated frame by frame by using each HMM. Such accumulated likelihood is calculated for every reference language, and the best reference language which has the largest accumulated likelihood is determined to be the uttered language. This procedure, illustrated in Figure 2, is said to capture the spectral differences that are present among languages.

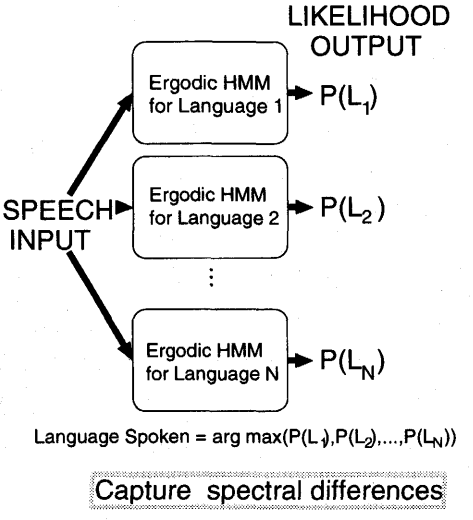

Figure 2: Ergodic HMM-based method

2.2 Trigram for Optimal State Sequence[14]

If each language characteristic takes account of phonotactics, we can consider the utilization of the optimal state sequence through the Viterbi algorithm. This fact was confirmed by a previous paper on a text database[15, 16]. Because each state of the HMM corresponds to vowel, consonant, silence, etc. respectively, we employ the 2nd order Markov model(trigram) to deal with an extracted optimal state sequence. First, we create an ergodic continuous density HMM using all language training data sets, and then produce optimal state sequences extracted by the Viterbi algorithm using this HMM for each language training data. The optimal state sequence is smoothed such that there will be no irregular state in the sequence, and then compressed so that the same successive states are united into one. However, in

order to consider the duration, if the length of the same successive states is long, the part is divided into several parts. This process is shown in Figure 3.

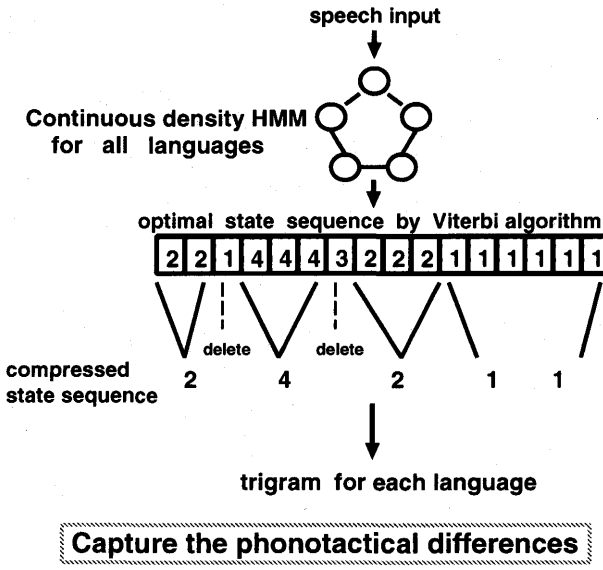

Figure 3: Optimal State Sequence Method

In our experiments, 2 or less same successive states are deleted and 3 or more same successive states are united into one every 100ms. These compressed optimal state sequences are used to create the trigram for each language.

The testing procedure produces the optimal state sequence obtained by the Viterbi algorithm using the universal HMM which is common to all languages. The likelihood of optimal state sequence is calculated by accumulating the probability for each trio of states while shifting one by one. Such accumulated likelihood is calculated for every reference language. This method is thought to capture the phonotactical differences among languages.

3 SPEECH MATERIALS AND ANALYSIS

3.1 Speech Database

The OGI Multi-Language Telephone Speech Corpus[1] contains utterances spoken by native speakers in 10 different languages that were collected over telephone lines. The ten languages are English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. The database consists of three sections: the training set(utterances from 50 speakers for each language), the development test set(utterances from 20 speakers for each language) and the final test set(utterances from 20 speakers for each language). Each speaker belongs only to one of the above sets. For this paper, we focused only on utterances from male speakers and excluded speech containing fixed vocabularies. The results reported

here were obtained by training with the training set and testing with the development test set. The training set has a total of 341 male speakers while the development test set has 151 male speakers.

3.2 Speech Analysis

OGI telephone speech are sampled at 8 kHz. The 14 cepstrum coefficients were calculated by the 14th order LPC analysis at a frame rate of 10ms with a window of 32ms. Then, 10 mel-cepstrum coefficients were obtained from the cepstrum coefficients. We also used regressive coefficients for each cepstrum time sequence as dynamic features, that is, 10 regressive coefficients (delta cepstrum). In this experiment, we did not use the prosodic information such as power, pitch and pause.

4 EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Test Data

As test data, only utterances from the male speakers of the development test set were used. Also, fixed vocabulary responses were excluded. While the utterances lasted an average of 13.6 seconds, the shortest utterance was 0.8 seconds with the longest at 49.4 seconds. Roughly, 26.7% of the utterances have a duration of over 10 seconds. The evaluation were performed for each of the test utterances.

4.2 OGI Ten Language Experiments

Ten language identification experiments were performed using the ergodic HMM and the trigram of optimal state sequence. The feature parameters of the ergodic HMM used here are mel-cepstrum coefficients and their regression coefficients. We should note here that the ergodic HMM used here has 10 states with one mixture for each state. Table 1 shows the identification results of the various conditions of ergodic HMMs and trigrams. Figure 4 shows the identification rates for the ergodic HMM, the trigram of optimal state sequence and their combination. The combination, which was realized by combining the likelihood from the ergodic HMM and the trigram, gave the best performance of 81.3% for the training data and 47.8% for the test data. The confusion matrices for the training data and test data are shown in Table 2(a),(b),respectively.

Table 1: Identification Rates for Ergodic HMM and Trigram[%]

	Ergodic HMM			Trigram	
	5 states	7 states	10 states	10 states	12 states
Training Data		69.1	72.5	58.4	
Test Data	36.1	38.5	36.6	39.9	38.8

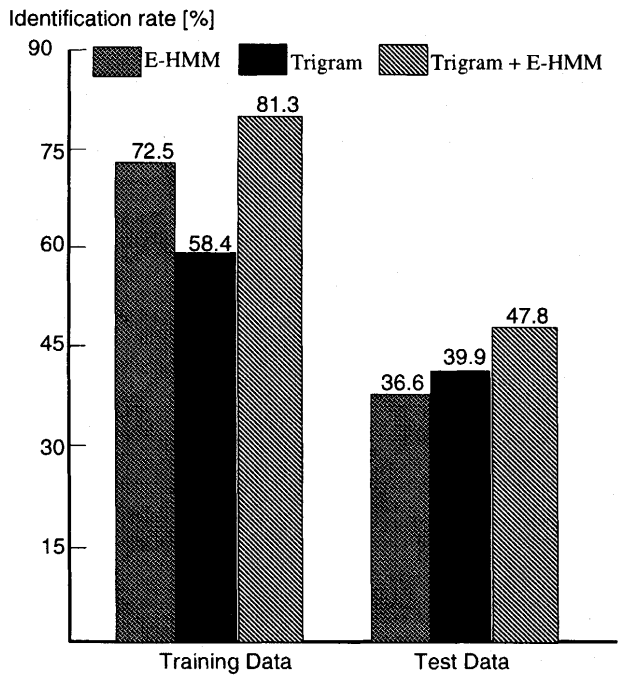

Figure 4: Ten Language Identification Results by Combining Ergodic HMM and Trigram of Optimal State Sequence

Table 2: 10-language Confusion Matrix for OGI Corpus

(a) Training Data

	Language Identified									
Input	E	Fa	Fr	G	J	K	M	S	T	V
En	169	3	0	1	3	2	2	3	1	0
Fa	7	156	4	8	2	8	12	2	5	4
Fr	4	4	192	3	4	6	2	5	2	3
Ge	4	1	3	117	0	2	1	3	3	3
Ja	5	1	3	1	136	7	5	2	3	4
Ko	0	2	4	1	7	132	9	2	0	8
Ma	5	3	7	2	3	4	148	0	2	7
Sp	4	9	4	1	4	6	3	151	5	4
Ta	7	5	1	1	4	10	5	15	176	7
Vi	3	4	2	1	6	4	4	0	9	126

(b) Test Data

	Language Identified									
Input	E	Fa	Fr	G	J	K	M	S	T	V
En	36	2	1	7	4	6	7	5	3	9
Fa	5	32	10	2	11	11	5	2	0	5
Fr	3	2	56	1	2	5	4	8	3	2
Ge	24	9	5	21	0	1	0	4	0	1
Ja	4	1	9	5	45	3	3	8	0	7
Ko	5	1	2	1	20	53	8	1	1	7
Ma	1	9	7	2	6	2	38	3	3	4
Sp	13	8	7	1	13	9	7	28	1	1
Ta	1	7	5	1	5	3	2	7	54	12
Vi	2	6	3	0	10	13	4	0	8	40

Since data from the OGI Telephone Speech Corpus were collected through telephone lines, a normalization procedure shown below was performed to reduce the effects of the communication channels.

$$\hat{x}_i = x_i - \bar{x} \tag{12}$$

In this procedure, the i-th frame of the normalized data $\hat{x}_i$ is obtained by substracting the mean $\bar{x}$ from the original data x_i. The mean $\bar{x}$ is the long-term cepstrum average per speaker for the training data and is the long-term cepstrum average per utterance for the test data.

After performing the above normalization procedure, identification rates have risen for all three methods as shown in Figure 5 (10-state ergodic HMM for each language and 10-state ergodic HMM for obtaining the optimal state sequences). Again, the combination of the ergodic HMM and the trigram gave the best identification rate of 57.5%, which represents a significant increase over the previous performance of 47.8%. The confusion matrices of the training data and test data for this combination are shown in Table 3 (a),(b). We can find that there are many confusable errors between sister languages, such as English and German, French and Spanish, Japanese and Korean, Tamil and Vietnamese and so on. Also, we find that Japanese and Spanish are confusable because their phoneme sequences consist of repeating vowel-consonant-vowel-consonant structure. These are marked by circles in Tables 3(b) and 5(b).

Furthermore, ergodic HMMs with one state and several mixtures were constructed and subsequently tested. The Gaussian mixture density functions are defined as follows:

$$P(x) = \Sigma_{k=1}^{K} \lambda_k \frac{1}{(2\pi)^n |\Sigma_k|^{\frac{1}{2}}} exp[-\frac{1}{2}(x-\mu_k)\Sigma_k^{-1}(x-\mu_k)^t], \tag{13}$$

where λ_k is the mixture weight, μ_k is the mean vector and Σ_k is the covariance matrix.

Test results for both training data and test data are shown in Table 4, where feature parameters were normalized for channel variations. From Table 4, we can see that the best performance was obtained with the model with 64 mixtures. Combining this model with the trigram of optimal state sequence, the identification rates for the training and test data were 89.0% and 58.5% , respectively, which equals the best performance reported in [10] for the OGI corpus. The confusion matrices for the combination are shown in Table 5. However, there is still a significant gap between the identification rates for the training data and test data and lessening this gap involves the consideration of more robust parameters.

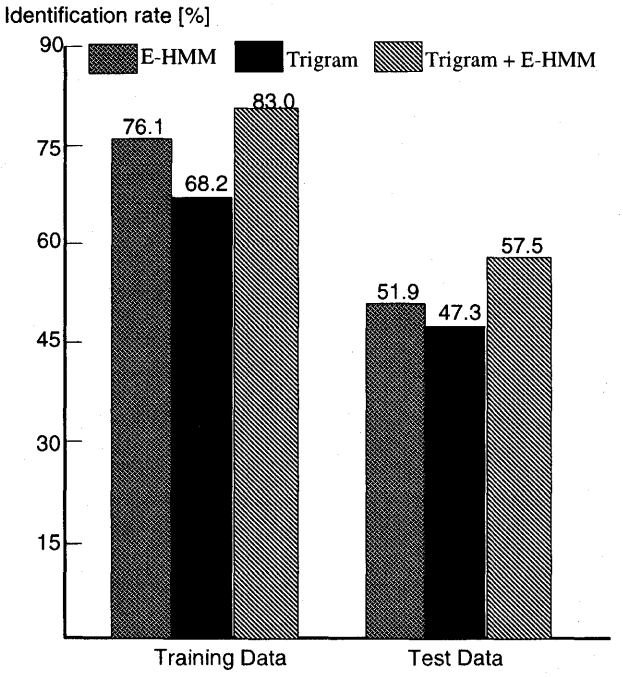

Figure 5: Identification results after Normalization

Table 3: Confusion Matrix after Normalization

(a) Training Data

	Language Identified									
Input	E	Fa	Fr	G	J	K	M	S	T	V
En	165	3	2	3	4	2	0	3	1	1
Fa	3	181	2	5	2	5	1	4	3	2
Fr	0	8	184	6	4	8	3	9	3	0
Ge	2	1	0	121	6	1	1	4	0	1
Ja	7	1	5	2	127	8	4	7	0	6
Ko	0	5	2	0	6	138	8	0	0	6
Ma	1	8	5	0	3	12	145	2	1	4
Sp	3	2	4	1	8	4	1	160	5	3
Ta	1	9	2	1	9	11	0	5	185	8
Vi	3	3	1	2	2	9	1	3	8	127

(b) Test Data

	Language Identified									
Input	E	Fa	Fr	G	J	K	M	S	T	V
En	49	5	1	4	2	2	1	5	6	5
Fa	2	48	6	4	7	8	1	2	0	5
Fr	1	0	50	2	3	9	5	⑯	0	0
Ge	⑬	2	6	27	4	0	0	13	0	0
Ja	2	0	2	1	64	3	1	⑨	1	2
Ko	0	1	3	1	⑯	72	1	3	0	2
Ma	5	3	5	1	3	9	28	10	2	9
Sp	1	5	3	3	⑱	11	1	38	2	6
Ta	1	6	3	0	3	9	1	5	59	⑩
Vi	1	2	2	0	6	9	5	0	⑪	50

Table 4: Identification results for the Gaussian Mixtures[%]

(a) Full covariance

Mixtures	Features	Training Data	Test Data
8	Ceps.	64.9	40.6
8	Ceps.+Delta Ceps.	67.5	47.0
16	Ceps.	72.8	41.2
16	Ceps.+Delta Ceps.	74.2	49.8
32	Ceps.+Delta Ceps.	80.2	46.1
64	Ceps.+Delta Ceps.	87.9	50.0

(b) Diagonal covariance

Mixtures	Features	Training Data	Test Data
64	Ceps.+Delta Ceps.	78.8	46.9

Table 5: Confusion Matrix for the Gaussian Mixture Density and Trigram Combination

(a) Training Data

	Language Identified									
Input	E	Fa	Fr	G	J	K	M	S	T	V
En	173	2	5	0	1	1	1	0	0	1
Fa	4	191	1	4	2	2	0	3	0	1
Fr	1	3	205	0	3	7	0	4	2	0
Ge	3	0	2	120	4	2	0	2	1	3
Ja	5	3	1	0	147	3	2	2	1	3
Ko	0	6	2	1	0	145	3	2	0	6
Ma	2	11	4	1	3	5	152	2	0	1
Sp	3	3	0	0	4	0	1	176	1	3
Ta	3	5	1	0	0	6	1	5	206	4
Vi	5	1	0	1	5	7	0	3	7	130

(b) Test Data

	Language Identified									
Input	E	Fa	Fr	G	J	K	M	S	T	V
En	50	6	0	2	2	2	1	5	4	8
Fa	3	47	7	6	5	6	1	1	0	7
Fr	4	0	52	2	1	8	5	⑬	1	0
Ge	⑯	3	10	21	4	1	0	10	0	0
Ja	1	0	4	1	60	5	2	⑨	1	2
Ko	1	2	1	0	⑯	70	3	3	1	2
Ma	3	6	6	0	1	6	41	7	2	3
Sp	1	10	⑨	4	⑮	6	1	34	3	5
Ta	0	4	0	2	1	7	1	8	63	⑪
Vi	0	4	3	0	6	8	4	1	4	56

5 SUMMARY

We performed a 10-language identification experiment using the OGI multi-language telephone speech corpus. Best performances were 89.0% and 58.5% for the training data and test data respectively, and these were accomplished by combining the ergodic HMM and the trigram of optimal sequence methods. While the results are among the best reported for the OGI corpus, the difference in the performance for the training and test data shows that more robust parameters have to be considered to lessen this gap.

Acknowledgement

The authors would like to thank Mr. Seino Takashi for his help in the development of the training and test programs and Mr. Konstantin Markov for his aid in the experiments using Gaussian mixtures.

References

[1] Y.K.Muthusamy, R.A.Cole, B.T.Oshika : The OGI Multi-Language Telephone Speech Corpus,Proc.ICSLP,pp.895-898(1992).

[2] P. Henrich, "Language Identification for the Automatic Grapheme-to-Phoneme Conversion of Foreign Words in a German Text-to-Speech System", Speech-89, pp.220-223(1989).

[3] A.House and E.Neuburg, "Toward Automatic Identification of the Language of an Utterance. I. Preliminary Methodological Considerations", J.Acous. Soc. Am., Vol.62, No.3, pp.708-713 (1977).

[4] R.Cole,et al., "Language Identification with Neural Networks: a Feasibility Study", Proc. IEEE, Pacific Rim Conference on Communications", pp.525-529 (1989).

[5] Y.K.Muthusamy and R.A.Cole, "Automatic Segmentation and Identification of Languages using Telephone Speech", Proc. ICSLP, pp.1007-1010 (1992).

[6] M.Sugiyama, "Automatic Language Identification Using Acoustic Features", Proc. ICASSP, pp.813-816 (1991).

[7] M.Savic, E.Acosta and S.K.Gupta, "An Automatic Language Identification System", Proc.ICASSP, pp.817-820(1991).

[8] M.A.Zissman, "Automatic Language Identification Using Gaussian Mixture and Hidden Markov Models", Proc. ICASSP, pp.II-399-402 (1993).

[9] M.A.Zissman, "Automatic Language Identification of Telephone Speech Messages Using Phoneme Recognition and N-Gram Modeling", Proc.ICASSP, pp.I-305-308 (1994).

[10] T.Hazen,V.W.Zue : Recent Improvements in an Approach to Segment-Based Automatic Language Identification, Proc.ICSLP, pp.1883-1886(1994).

[11] S.Nakagawa, Y.Ueda and T.Seino , "Speaker -independent, Text-independent Language Identification by HMM", Proc. ICSLP, pp.1011-1014 (1992).

[12] L.F.Lamel and J.L.Gauvain, " Language Identification Using Phone-based Acoustic Likelihoods", Proc. ICASSP, pp.293-296 (1994).

[13] A.A.Reyes, T.Seino and S. Nakagawa, "Three Language Identification Methods Based on HMMs", to be published in Proc. ICSLP (1994).

[14] S.Nakagawa and T.Seino, "Spoken Language Identification Using Ergodic HMM with Emphasized State Transition", Proc. Eurospeech, pp.133-135(1993).

[15] Y.Ueda and S.Nakagawa, "Prediction for Phoneme / Syllable / Word Category and Identification of Language Using HMM", Proc. ICSLP, pp.1209-1212 (1990).

[16] S.Nakagawa and Y.Ueda, "Automatic Extraction of Phonotactics Based on Hidden Markov Models and Language Identification", Studia Phonologica, XXIV, pp.83-95(1990).