PAPER *Special Issue on Spoken Language Processing*

Automatic Language Identification Using Sequential Information of Phonemes

Takayuki ARAI[†], ***Member***

SUMMARY In this paper approaches to language identification based on the sequential information of phonemes are described. These approaches assume that each language can be identified from its own phoneme structure, or phonotactics. To extract this phoneme structure, we use phoneme classifiers and grammars for each language. The phoneme classifier for each language is implemented as a multi-layer perceptron trained on quasi-phonetic hand-labeled transcriptions. After training the phoneme classifiers, the grammars for each language are calculated as a set of transition probabilities for each phoneme pair. Because of the interest in automatic language identification for worldwide voice communication, we decided to use telephone speech for this study. The data for this study were drawn from the OGI (Oregon Graduate Institute)-TS (telephone speech) corpus, a standard corpus for this type of research. To investigate the basic issues of this approach, two languages, Japanese and English, were selected. The language classification algorithms are based on Viterbi search constrained by a bigram grammar and by minimum and maximum durations. Using a phoneme classifier trained only on English phonemes, we achieved 81.1% accuracy. We achieved 79.3% accuracy using a phoneme classifier trained on Japanese phonemes. Using both the English and the Japanese phoneme classifiers together, we obtained our best result: 83.3%. Our results were comparable to those obtained by other methods such as that based on the hidden Markov model.

key words: *language identification, phonetic transcription, telephone speech, multi-layer perceptron, bigram*

1. Introduction

Automatic language identification can be very useful in worldwide voice communication networks. A telephone interpreter (automatic or a group of humans), for example, needs a way to identify the language being spoken. A language identifier could then switch to an operator who can speak that language or to an appropriate speech recognition system.

Previous work has combined the use of inherent and prosodic features[1]–[4]. Sugiyama[5] has relied on long utterances to build up representative histograms of acoustic feature vectors before classifying the specific language spoken. Recently, ergodic HMMs (hidden Markov models) have been used by Nakagawa and other researchers to obtain this information[6]–[10]. Different from such frame-based approaches, phone-based or segment-based approaches are used[9], [11]. A segmental approach, which assumes that the acoustic structure of languages can be estimated by segmenting speech into phonetic categories, has been applied by Muthusamy using broad phonetic categories[12], [13]. Zissman has achieved higher performance[14] using an N-gram with phoneme recognition. Berkling[15] showed that mono- and poly-phonemes contain useful information.

Because of the interest in automatic language identification for worldwide communication networks, we decided to concentrate on telephone speech in this study. The data for this study were drawn from the OGI (Oregon Graduate Institute)-TS (telephone speech) corpus, which is a standard corpus for this type of research. Some other researchers are already using this corpus[9], [11], [14]–[16].

The underlying hypothesis of the approach detailed in this paper is that each language has its own phoneme structure, or phonotactics. House[17] suggested that an approach based on phonotactics is useful. If one can observe the sequence of phonemes, one can classify the language easily. To extract this phoneme structure, we use phoneme classifiers and grammars for each language.

In our approach, we used first-order Markov models, that is, bigrams or phoneme pairs to handle sequential information. The phoneme classifier for each language is implemented as a multi-layer perceptron trained on quasi-phonetic hand-labeled transcriptions. After training the phoneme classifiers, from the grammar of each language we calculate a set of transition probabilities for each phoneme pair. To investigate the basic issues of this approach, two languages, Japanese and English, were selected. In this paper, several algorithms for language identification of speaker-independent and text-independent language based on telephone speech are discussed. In Sect. 2, our speech corpus and Japanese transcription are described. We developed conventions for Japanese quasi-phonetic transcription and added them to the original conventions for English. In Sect. 3, we describe the phoneme classifier, which is the first stage of the system. In Sect. 4, we describe two approaches to classifying the two languages. The first result is obtained using the Viterbi score from a front-end of each language and the second from a combination of these two front-ends.

Manuscript received October 27, 1994.

Manuscript revised January 31, 1995.

[†] Department of Electrical and Electronic Engineering, Sophia University, Tokyo, 102 Japan.

Table 1 Number of callers classified by gender for each data set in OGI-TS corpus.

Language	Training		Development Test		Final Test *	
	Males	Females	Males	Females	Males	Females
English	33	17	14	6	16	4
Japanese	31	19	15	5	10	9

* There was a Japanese caller whose gender was unknown.

Table 2 Six utterances for each speaker.

htl	Something that he/she likes about his/her hometown. (10 s)
htc	About the climate in his/her hometown. (10 s)
room	Description of the room that he/she is calling from. (12 s)
meal	Description of his/her most recent meal. (10 s)
story-bt	First 45 seconds of one-minute free speech before the tone. (45 s)
story-at	Last 10 seconds of one-minute free speech after the tone. (10 s)

2. Speech Data

2.1 Speech Corpus

Continuous telephone speech in English and Japanese from the multi-language OGI-TS corpus[18] was used for training and testing. The speech data were sampled at 8 kHz with a 13-bit resolution. For each language, this corpus contains 50 calls for training, 20 calls for a development test, and 20 calls for an evaluation test. Each call was uttered by a unique adult speaker whose of gender is shown in Table 1. From the speech of each person, the six utterances in Table 2 including the topic-specific descriptions and a one-minute free speech were chosen. The one-minute free speech was divided into two utterances before and after the tone, which informed the caller that the remaining time was ten seconds.

The utterances of the training data set are labeled in two stages, word labeling and quasi-phonetic transcription, to train phoneme classifiers based on neural networks. Using two stages in the automatic forced phoneme alignment by word labeling helps to streamline the transcription of the phonemes. The quasi-phonetic transcription was done by native-speaker experts, who have knowledge of phonetics and expertise of labeling, using waveform and spectrogram as shown in Fig. 1.

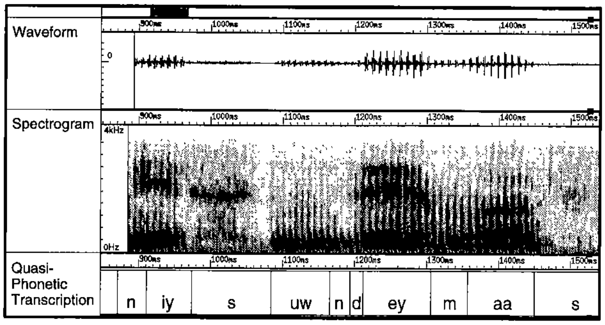


Fig. 1 Waveform and spectrogram for the quasi-phonetic transcription.

2.2 Japanese Phonetic Transcription

After Japanese utterances in the training data set are labeled at the word level using a Japanese romanization, the utterances are labeled at the phonemic level using a quasi-phonetic transcription. The quasi-phonetic transcription does not include all phonetic information; some phonetic details are added to the phonemic labeling scheme. This quasi-phonetic transcription is based on Japanese phonetics[19].

The symbols for the five Japanese vowels are shown in Table 3. These are nominally a subset of English vowels, but the pronunciation is different. In particular, /ow/ and /ey/ are not diphthongs as in English. These symbols are used because the initial vowel is close to the Japanese vowel. So the symbols we will use are these: /aa/ for [ɑ], /ey/ for [e], /iy/ for [i], /ow/ for [o], and /uw/ for [ɯ]. The bracketed characters are those used in the IPA (International Phonetic Alphabet).

Table 3 Symbols used for quasi-phonetic transcription of the five Japanese vowels.

Quasi-Phonetic Symbol	Tongue-height	Tongue-position	Aperture
/iy/	high	front	unrounded
/ey/	mid	front	unrounded
/aa/	low	central	unrounded
/ow/	mid	back	rounded
/uw/	high	back	unrounded

The mapping from acoustic sounds of Japanese consonants to symbols which we are using is shown in Table 4. Phonetically Japanese has more sounds, but some allophones are represented with one symbol. For example, both the common voiceless bilabial fricative

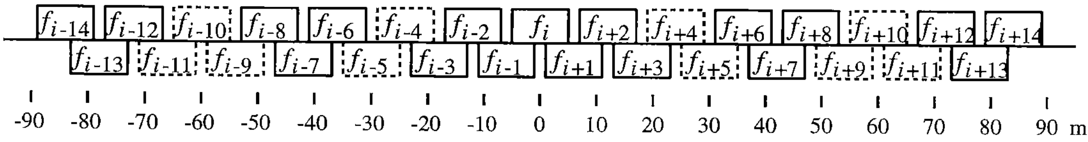


Fig. 2 Seven averaged feature vectors computed for sampled frames (solid boxes).

Table 4 Correspondence between quasi-phonetic transcription and IPA for Japanese consonants.

Quasi-Phonetic Symbol	International Phonetic Alphabet	Quasi-Phonetic Symbol	International Phonetic Alphabet
/p/	[p]	/sh/	[ʃ]
/t/	[t]	/hh/	[h],[ɦ],[ç]
/k/	[k]	/ts/	[ts]
/b/	[b]	/ch/	[ʧ]
/d/	[d]	/dz/	[dz],[z]
/g/	[g]	/jh/	[ʤ],[ʒ]
/m/	[m]	/rx/	[ɾ],[r],[l]
/n/	[n],[ɲ],[ŋ],[N]	/w/	[w]
/f/	[f],[ɸ]	/y/	[j]
/s/	[s]		

Table 5 Division of frames for phoneme classifier into training and test sets.

	Train	Test
	Utterances (Frames)	Utterances (Frames)
English Phonetic Classifier	50 (80502)	20 (7441)
Japanese Phonetic Classifier	45 (23546)	10 (5608)

[ɸ] and the rarer voiceless labio-dental fricative [f] are represented with /f/. Likewise, most /rx/ are the flap [ɾ] while some are the voiced retroflex stop [ɖ] at the word-initial position or liquid-like [r],[l] in rapid or emphasized speech. Some /hh/ sounds are very close to the German "ich-Laut" [ç]. There are also palatal nasals [ɲ], velar nasals [ŋ] and uvular nasals [N] in Japanese, but sometimes these are allophones and labeled as /n/. In addition, the voiced alveolar fricative [z] and affricate [dz] are allophones and are labeled as /dz/; voiced palato-alveolar fricative [ʒ] and affricate [ʤ] are allophones and are labeled as /jh/. The English symbol set has a voiceless palato-alveolar affricate [ʧ] but not a voiceless alveolar affricate [ts], which is added for Japanese.

3. Phoneme Classification

This section describes a phoneme classification by an MLP (multi-layer perceptron) of the system. Features are derived from the signal in order to perform a frame-based phoneme classification of the incoming speech signal. Frame-based phoneme classification provides a score for a set of phonemic labels which can approximate the probability of a phoneme occurring at the input level.

3.1 Feature Generation

A seventh-order Perceptual Linear Predictive (PLP) model yielding 8 coefficients (including one for energy) was computed on a 10 ms interval of speech, time-shifted every 6 ms. The technique of PLP uses basic concepts from the psychophysics of hearing while maintaining a low dimensional representation of speech [22]. This approach has been proven to be useful in speaker-independent automatic speech recognition [20].

For the ith frame with a feature vector f_i, seven averaged feature vectors were computed as follows:

$$\begin{aligned}&(f_{i-14}+f_{i-13}+f_{i-12})/3,\\&(f_{i-8}+f_{i-7}+f_{i-6})/3,\\&(f_{i-3}+f_{i-2}+f_{i-1})/3,\\&f_i,\\&(f_{i+3}+f_{i+2}+f_{i+1})/3,\\&(f_{i+8}+f_{i+7}+f_{i+6})/3,\\&(f_{i+14}+f_{i+13}+f_{i+12})/3.\end{aligned}$$

The seven averaged feature vectors are taken from 7 regions spanning a 178 ms window centered on the frame f_i to be classified as shown in Fig. 2. The solid boxes are sampled and the dashed boxes are skipped. These features were empirically derived to capture the contextual information in the vicinity of each frame [21]. The objective was to provide substantial contextual information about the chosen frame to the network.

Seven averaged feature vectors (56 averaged PLP coefficients) which include seventh order PLP coefficients and energy are given as input to each of the phoneme classifiers [23]. The PLP coefficients are normalized between -1 and 1 to decrease the probability that the MLP will converge to an undesirable local minimum. Due to computational complexity, the system was trained from a randomly chosen subset of the frames from the training utterances shown in Table 5.

3.2 Training for MLP

For both languages, classification is performed by a fully-connected, three-layer, feedforward network that assigns phonemic category scores to each 6 ms time-frame. The labels provide a quasi-phonetic level of de-

Table 6 Division of English and Japanese utterances for language classifier into training and test set. (The average length of the utterances is 13.5 seconds.)

	Train		Test	
	Utterances	Avg. [s]	Utterances	Avg. [s]
English Language Classifier	279	13.7	115	13.6
Japanese Language Classifier	278	13.5	112	13.3

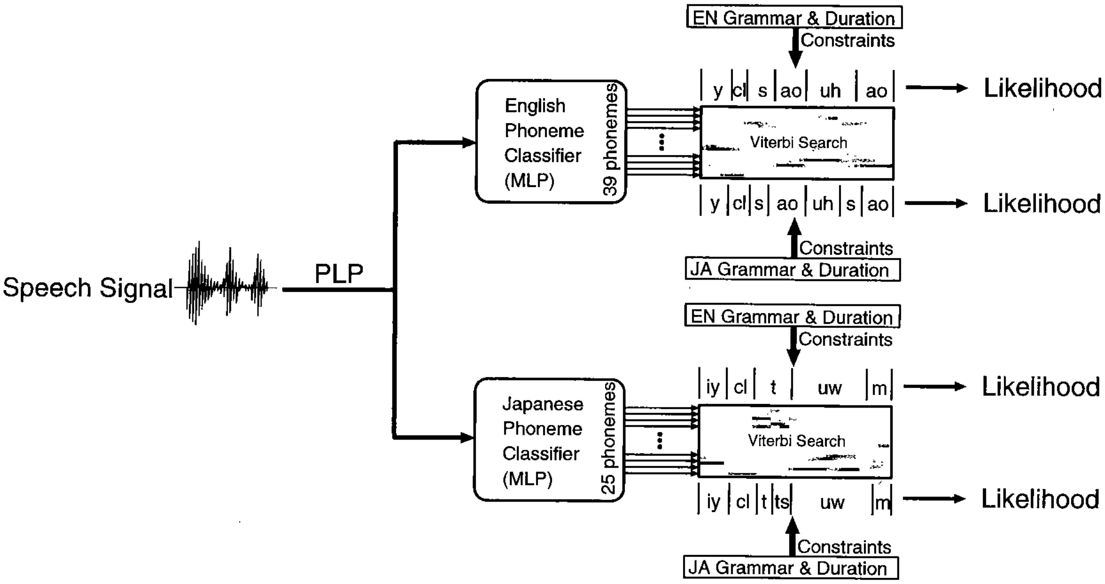


Fig. 3 Block diagram of the system including an MLP (multi-layer perceptron) and Viterbi search.

scription in which most allophonic variations are ignored. Some similar categories are combined: /ih/-/ix/, /ah/-/ax/, /aa/-/ao/, /m/-/em/, /n/-/en/-/ng/-/nx/, /sh/-/zh/, and /hh/-/hv/ in English. These labels are a slightly modified version of the labels used in the public-domain TIMIT data base[24]. Finally, 39 categories including one for closure are used for English phonemes:

/cl/, /iy/, /ih/, /ey/, /ae/, /eh/,/ah/,/uw/,
/uh/,/ow/,/aw/,/aa/,/ay/,/oy/,/er/, /ao-r/,
/r/, /l/, /y/, /w/, /hh/,/m/,/n/, /s/,
/z/, /sh/, /th/, /dh/,/dx/,/f/, /v/, /ch/,
/jh/,/p/, /t/, /k/, /b/, /d/, /g/.

The English network has 56 input nodes, 48 hidden nodes, and 39 output nodes, one for each of the English phonemes. For Japanese phonemes, 25 categories including one for closure are used:

/cl/,/iy/, /ey/,/uw/,/ow/,/aa/,/rx/,/y/,
/w/,/hh/,/m/,/n/, /s/, /sh/,/f/, /ch/,
/jh/,/p/, /t/, /k/, /b/, /d/, /g/, /dz/,
/ts/.

The Japanese network has 56 input nodes, 48 hidden nodes, and 25 output nodes, one for each of the Japanese phonemes. The number of hidden nodes was determined experimentally.

The networks were trained using back-propagation with conjugate gradient optimization[25]. Training continued until the performance of the network on the test set peaked. Classification performance was evaluated on the frames in the test set given in Table 5. The phoneme classifier performed with a frame-by-frame accuracy of 48.0% for English and 44.0% for Japanese.

4. Language Classification

4.1 Training for Language Identification

We tried two approaches to language classification, one comparing the Viterbi score of each language, and another using both classifiers. In each case, the training and development test data in Table 6 was used. Figure 3 shows an example of the sequence of phonemes after the Viterbi search.

The English and Japanese networks was both used for input from either language and the likelihood from the Viterbi search for each language were compared. The frame-by-frame outputs of the phoneme classifier were converted into a time-aligned sequence of phoneme labels using a Viterbi search which maximizes

$$L_t(j) = \max_i \left[L_{t-1}(i) + \log[P(p_j|p_i)] + \alpha(d) \right] + \log[o_t(j)], \quad (1)$$

where $L_t(j)$ is the logarithmic likelihood of phoneme p_j at frame t, $P(p_j|p_i)$ is the transition probability from phoneme p_i to p_j, and $o_t(j)$ is the activation of phoneme p_j from the phoneme classifier at frame t. The initial condition is given by

$$L_0(i) = \log \pi(i). \quad (2)$$

For this experiment, $\pi(i)$ were set to 1.0 for all i. To control the duration of phonemes, $\alpha(d)$ was added to the Eq. (1) to impose a penalty. The $\alpha(d)$ is a function of the duration d of the current phoneme p_j as follows:

$$\alpha(d) = \begin{cases} 0 & d_{\min}(j) \leqq d \leqq d_{\max}(j), \\ A & \text{otherwise,} \end{cases} \quad (3)$$

where $d_{\min}(j)$ and $d_{\max}(j)$ are the minimum and maximum duration for the phoneme p_j and A is a negative number. For each language, the grammar, or transition probability $P(p_j|p_i)$ for all i and j was estimated from the training sets. To get the transition probability $P(p_j|p_i)$, a Viterbi search of the training data phoneme outputs was performed under the condition that the durations for that training data set be reasonable. We allowed the duration to be from 3 to 300 frames for all phonemes ($d_{\min} = 3$, $d_{\max} = 300$); these number were obtained empirically. We employed the sequence of phonemes from all training utterances.

4.2 Results for Language Identification

We evaluated our system in two steps. We first report results from the Viterbi score of each phoneme classifier, then we report results using both classifiers. We also compared our results with reported results which used other techniques.

For testing the performance, a second Viterbi search was done under the condition set by the bigram using $P(p_j|p_i)$ for each language. In a preliminary experiment, the performance of the language identification with $d_{\min}(j)$ and $d_{\max}(j)$ for each phoneme p_j obtained from all training utterances was 0.4% higher than that with the fixed $d_{\min}(j)$ and $d_{\max}(j)$ for all phonemes. We confirmed that the minimum and maximum duration influence the accuracy very little. Therefore, the condition of duration was the same as for the training data set, that is, the allowed duration was from 3 to 300 frames for all phonemes ($d_{\min} = 3$, $d_{\max} = 300$). The difference between the number of categories in English and that in Japanese introduces a bias for each language. In practice, the logarithmic likelihood was scaled by a factor β, chosen to give optimal training-set performance, in order to reduce the influence. Finally, likelihood L was given by

$$L = \beta \max_j L_T(j), \quad (4)$$

where T indicates the final frame. The factor β was 0.99813 for an English phoneme classifier and 1.00360 for a Japanese phoneme classifier. The classifier assigned an incoming utterance to the language with the largest likelihood according to this expression.

The results using the Viterbi score are shown in Table 7. When we used an English phoneme classifier, we obtained 81.1% of accuracy. When we used a Japanese phoneme classifier, we obtained 79.3% of accuracy. After combining those two outputs of likelihood from both classifiers, we obtained 83.3% of accuracy. This shows that the combination of the two classifiers has a better performance because the system works with more information.

To compare our approach to that of others, we used the same OGI-TS corpus. Two other approaches were tested and compared to ours; the corpus and results are shown in Table 8. One other approach (by Muthusamy [12]) is based on a broad phonetic category using both statistical and sequential features. The results of our approach compare favorably to the 83.6% recognition rate achieved with the phonemic approach using bigram information. The second approach tested is HMM, which recently is commonly used for speech recognition and other speech processing. The method of Zissman [14], the results based on the HMM using Gaussian mixture were 82.9% for 45-second story-bt utterances. Under the same conditions, for story-bt utter-

Table 7 Results of language classification between English and Japanese using Viterbi score.

Phoneme Classifier	Accuracy
English	81.1%
Japanese	79.3%
English and Japanese	83.3%

Table 8 Proposed method compared to others.

Approach	Corpus	Accuracy
Our method	whole set	83.3%
Broad phonetic category *	whole set	83.6%
Our method	story-bt	88.6%
HMM using Gaussian mixture * *	story-bt	82.9%

* Muthusamy [13]
* * Zissman [14]

ances our method gave comparable results. For the same utterances we obtained 88.6% which was 5.3% higher than what we obtained for the whole corpus. This was probably caused by the difference of the average length of those two sets of the utterances, that is, 13.5 seconds for the whole corpus and 48.4 seconds for the story-bt utterances.

5. Conclusion

Some approaches for language identification based on the sequential information of phonemes have been described. We focused on telephone speech in two languages, Japanese and English, from OGI-TS corpus. This corpus and our Japanese transcription were also described. The phoneme classifier for each language is implemented as an MLP, trained on quasi-phonetic hand-labeled transcriptions. The phoneme classifier performed with a frame-by-frame accuracy of 48.0% for English and 44.0% for Japanese. The performance is not high enough and it was probably due to the highly variable conditions, especially the telephone line and the variety of speakers. Zissman [7] showed that the cancellation of the channel effect was highly useful. The improvement of the performance of the phoneme classifiers is the next problem.

After training the phoneme classifiers, a set of transition probabilities for each phoneme pair are calculated for each language. Experimental results were obtained using the Viterbi score from separate front-ends and from both front-ends using bigrams. The approach using the Viterbi score has low computational complexity because it only requires the likelihood from the Viterbi search. We obtained 83.3% as our best performance for the whole corpus and 88.6% for the story-bt corpus. Compared to other methods, our results were comparable.

In the future we intend to test these various approaches on other languages drawn from the ten-language OGI-TS corpus. The additional complexity needed to extend the number of languages to be identified, calls for a system with a simpler structure. For simplification perhaps only one phoneme classifier could be used for all languages. With more languages, we would like to consider using separate grammar and phoneme classifiers for each language.

Acknowledgements

I would like to acknowledge gratefully the assistance of the following persons. Dr. R. Cole gave me the opportunity to study this field; Dr. E. Barnard gave me valuable advice and suggestions for this research; Dr. B. Oshika helped develop the transcription conventions; Dr. H. Hermansky has developed the PLP analysis; Dr. M. Fanty wrote the test environment and speech recognition system that I used and he also gave me several helpful ideas. I am also indebted to Dr. Y. Muthusamy who implemented the original language identification research environment, modified the speech recognition system to perform language identification, and advised me on this research. I also want to thank Kay Berkling who worked together with me and helped in writing this paper, Karen Ward who edited the first draft of this paper, and the other members of the Center for Spoken Language Understanding (CSLU) at the Oregon Graduate Institute of Science & Technology (OGI).

References

[1] Foil, J.T., "Language identification using noisy speech," In *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, vol.2, pp.861–864, Apr. 1986.

[2] Goodman, F.J., Martin, A.F. and Wohlford, R.E., "Improved automatic language identification in noisy speech," In *Proc. Int'l Conf. of the American Society of Signal Processing*, pp.528–531, 1989.

[3] Savic, M., Acosta, E. and Gupta, S.K., "An automatic language identification system," In *Proc. Int'l Conf. of the American Society of Signal Processing*, pp.817–820, 1991.

[4] Muthusamy, Y.K. and Cole, R.A., "A segment-based automatic language identification system," In Moody, J.E., Hanson, S.J. and Lippmann, R.P., editors, *Advances in Neural Information Processing Systems*, vol.4, Morgan Kaufmann, 1992.

[5] Sugiyama, M., "Automatic language recognition using acoustic features," In *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, vol.2, pp.813–816, May 1991.

[6] Nakagawa, S., Ueda, Y. and Seino, T., "Speaker-independent, text-independent language identification by HMM," In *Proc. Int'l Conf. on Spoken Language Processing*, Oct. 1992.

[7] Zissman, M.A., "Automatic language identification using Gaussian mixtures and hidden Markov models," In *Proc. IEEE Int'l Conf. of Acoustics, Speech, and Signal Processing*, vol.2, pp.399–402, Apr. 1993.

[8] Seino, T. and Nakagawa, S., "Spoken language identification using ergodic HMM with emphasized state transition," In *Proc. European Conf. on Speech Communication and Technology*, pp.133–136, Sep. 1993.

[9] Lamel, L.F. and Gauvain, J.L., "Language identification using phone-based acoustic likelihoods," In *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, vol.1, pp.293–296, Apr. 1994.

[10] Tucker, R.C.F., Carey, M.J. and Parris, E.S., "Automatic language identification using sub-word models," In *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, vol.1, pp.301–304, Apr. 1994.

[11] Hazen, T.J. and Zue, V.W., "Automatic language identification using a segment-based approach," In *Proc. European Conf. on Speech Communication and Technology*, pp.1303–1306, Sep. 1993.

[12] Muthusamy, Y.K., Berkling, K.M., Arai, T., Cole, R.A. and Barnard, E., "A comparison of approaches to automatic language identification using telephone speech," In *Proc. of the European Conf. on Speech Communication and Technology*, pp.1307–1310, Sep. 1993.

[13] Muthusamy, Y.K., "A segmental approach to automatic language identification," Ph.D. dissertation, Oregon Graduate Institute of Science & Technology, Oregon, U.S.A., 1993.

[14] Zissman, M.A. and Singer, E., "Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling," In *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing,* vol.1, pp.305–308, Apr. 1994.

[15] Berkling, K.M., Arai, T. and Barnard, E., "Analysis of phoneme-based features for language identification," In *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing,* vol.1, pp.289–292, Apr. 1994.

[16] Berkling, K.M. and Barnard, E., "Language identification of six languages based on a common set of broad phonemes," In *Proc. Int'l Conf. on Spoken Language Processing,* vol.4, pp.1891–1894, Sep. 1994.

[17] House, A.S. and Neuberg, E.P., "Toward automatic identification of the language of an utterance. I. Priliminary methodological considerations," *Journal of the Acoustical Society of America,* vol.62, no.3, pp.708–713, 1977.

[18] Muthusamy, Y.K., Cole, R.A. and Oshika, B.T., "The OGI multi-language telephone speech corpus," In *Proc. Int'l Conf. on Spoken Language Processing,* Oct. 1992.

[19] Vance, T.J., *An Introduction to Japanese Phonology,* State University of New York Press, Albany, 1987.

[20] Creekmore ,J.W., Fanty, M.A. and Cole, R.A., "A comparative study of five spectral representations for speaker-independent phonetic recognition," In *Proc. 25th Annual Asilomar Conf. on Signals, Systems, and Computers,* 1991.

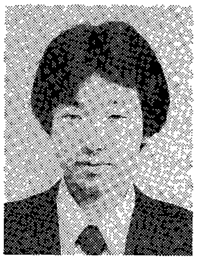
[21] Fanty, M.A., Cole, R.A. and Roginski, K., "English alphabet recognition with telephone speech," In Moody, J.E., Hanson, S.J. and Lippmann, R.P., editors, *Advances in Neural Information Processing Systems 4,* San Mateo, CA, Morgan Kaufmann Publishers, 1992.

[22] Hermansky, H., "Perceptual linear predictive PLP analysis of speech," *Journal of the Acoustical Society of America,* vol.4, pp.1738–1752, Apr. 1990.

[23] Fanty, M.A., Schmid, P. and Cole, R.A., "City name recognition over the telephone," In *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing,* vol.1, pp.549–552, Apr. 1993.

[24] Fisher, W.M., Doddington, G.R. and Goudie-Marshall, K.M., "The DARPA speech recognition research database: Specification and status," In *Proc. DARPA Speech Recognition Workshop,* pp.93–100, Feb. 1986.

[25] Barnard, E. and Cole, R.A., "A neural-net training program based on conjugate-gradient optimization," *Technical Report,* vol.CSE 89-014, Oregon Graduate Institute, 1989.



Takayuki Arai was born in Tokyo, Japan, on November 12, 1966. He received the B.E., M.E., and Ph.D. degrees all in electrical and electronic engineering, from Sophia University, Tokyo, Japan, in 1989, 1991 and 1994, respectively. From 1993 until 1994, he studied at the Oregon Graduate Institute of Science and Technology, U.S.A. Currently, he is a research assistant in the Department of Electrical and Electronic Engineering of Sophia University, engaged in education and research in the fields of speech signal processing.