# LANGUAGE IDENTIFICATION USING PHONETIC AND PROSODIC HMMS WITH FEATURE NORMALIZATION

*Yasunari Obuchi and Nobuo Sato*

Advanced Research Laboratory, Hitachi Ltd.
Kokubunji, Tokyo 185-8601, Japan
{obuchi, n-sato}@rd.hitachi.co.jp

## ABSTRACT

Phonetics and prosody are two important factors in automatic language identification. Prosodic HMMs enable language identification systems to use prosodic information in a similar manner to phonetic HMMs. This paper describes how to create prosodic HMMs and implement them in language identification systems. Linear discriminant analysis of the likelihood and N-gram scores of prosodic segment recognition realizes fast and reliable language identification. Moreover, combining prosodic HMMs with phonetic HMMs improves system performance. In this framework, feature normalization techniques that were originally developed for robust speech recognition can be applied to phonetic and prosodic features. Language identification accuracy increases using these techniques in clean and noisy environments.

## 1. INTRODUCTION

Automatic language identification (LID) systems identify the language using phonetic and/or prosodic information extracted from input speech. Recent successes in automatic speech recognition have established a way to analyze phonetic information using Hidden Markov Models (HMMs), which were also used in many LID systems. In contrast, prosodic information is used in a simple way, and the precise dynamics of prosody has been largely ignored. In this paper, we will introduce prosodic HMMs, and apply them to English, Japanese, and Mandarin Chinese LID systems in a similar manner to phonetic HMMs.

In the previous studies [1, 2], prosodic information was statically processed, and the frame or segment-based likelihood scores were simply accumulated over an utterance. In contrast, Adami et al. [3] proposed a new approach that tokenizes the prosodic segment. They classified prosodic segments using ad-hoc rules about the power and fundamental frequency (F0). In their work, the phonotactic approach was extended to prosody, taking inter-segmental dynamics into account. However, the intra-segmental dynamics has not yet been considered. Therefore, we will introduce prosodic HMMs for classifying the prosodic segments dynamically.

In the case of phonetic HMMs, parallel phone recognition (PPR) and parallel phone recognition followed by language modeling (PPRLM) [4] are two well-known algorithms. In this paper, we will show that prosodic HMM realizes the implementation of prosody-based LID analogous to PPR and PPRLM. Moreover, when the PPR- and PPRLM-based approaches for phonetic and prosodic features are combined, the system performs more effectively.

Another benefit of prosodic HMM is that feature normalization techniques for phonetic HMM can be applied to them. We will demonstrate that mean and variance normalization (MVN) and delta-cepstrum normalization (DCN) improve LID performance, especially under noisy conditions.

The remainder of this paper is organized as follows. In the next section, we will describe the creation of prosodic HMMs. Section 3 describes how to implement the combined LID system using phonetic and prosodic HMMs. In Section 4, we will describe feature normalization techniques applicable to both phonetic and prosodic features. Section 5 presents experimental results, and the final section gives conclusions and future works.

## 2. PROSODIC HMM

Iwano et al. [5] proposed prosodic HMM for robust speech recognition, where they used fixed categories defined by the F0 transition pattern. In this paper, we use a data-driven clustering technique to create language-dependent prosodic HMMs. The number of categories can be chosen arbitrarily. Since it is difficult to obtain multi-language corpora with prosodic labels, we created prosodic HMMs by unsupervised training. A similar approach was used by Nagarajan et al. [6] to create phonetic HMMs of syllable-like units. A more detailed description of HMM creation follows.

First, the training data of each language are segmented into frames. Common frame length and rate values are used in phonetic and prosodic processing. For each frame, the power, F0, and reliability of the F0 estimate form a feature

0-7803-8874-7/05/$20.00 ©2005 IEEE

vector. The cepstrum method is used to estimate F0, where the cepstrum peak and its intensity in the 40- to 500-Hz regions are regarded as F0 and its reliability. Therefore, F0 is set even in unvoiced regions, but its reliability has a small value. Next, these feature vectors are concatenated for 2N frames to form a larger feature vector of 6N elements. Here, N is the number of states in a prosodic HMM. It is assumed that the average duration of a state is two frames. Therefore, the elements of the two neighboring frames are averaged, resulting in 3N elements in a feature vector: power, F0, and F0 reliability of N states. Finally, the first and second order time-derivatives of these features are added to increase the dimension to 9N. These feature vectors are clustered to make initial HMMs using the K-means algorithm. The output probability of each state is expressed by a single Gaussian probability density function. The transition probabilities are set to 0.5 for self and next transitions.

After creating the initial HMMs, they are used to decode the training data, and the first transcription is obtained. The first transcription is used for the first run of Viterbi training to make the second HMMs, and these procedures are repeated for a pre-defined number of times. This is the end of the prosodic HMM creation for one language, and the same procedure is executed for other languages. As for the phonetic HMMs, standard training is carried out for each language using phonetically labeled corpora. We also make prosodic/phonetic N-gram language models by decoding the training data using prosodic/phonetic HMMs.

## 3. COMBINED LID SYSTEM

Prosodic HMMs and phonetic HMMs are combined in a simple manner as shown in Fig. 1. The LID system consists of a feature extraction module, three language-dependent scoring modules for English (ENG), Japanese (JPN), and Mandarin (MND), and a linear discriminant analysis (LDA) module. The feature extraction module extracts the MFCC and prosodic features (and their time derivatives), and sends them to the scoring modules. In the scoring module, the phone recognition unit converts the feature vectors into a phone sequence using phonetic HMMs, where a phonetic likelihood score is obtained at the same time. The phone sequence is evaluated by the phonetic N-gram to provide a phonetic N-gram score in the same manner as PPRLM in [4]. The prosody recognition unit works in a similar fashion, providing a prosodic likelihood score and a prosodic N-gram score. Thus, the three scoring modules give 12 scores in total, all of which are fed into the LDA module. This module makes the final decision using these scores.

There are various ways of making the final decision from a set of LID scores. Some examples of fusing algorithms are described in [7]. In this work, we applied a statistical fusion technique, based on linear discriminant analysis in two domains, phonetics and prosody. In each domain, any pair of languages are discriminated using both N-gram and likelihood scores. Five discriminant coefficients (for N-gram and likelihood scores of two languages, plus a threshold) are calculated for each pair using the development data, and a discriminant score is computed for a test utterance using these coefficients. Then, the discriminant score in the phonetic and prosodic domains are added, and a decision is made for the pair.

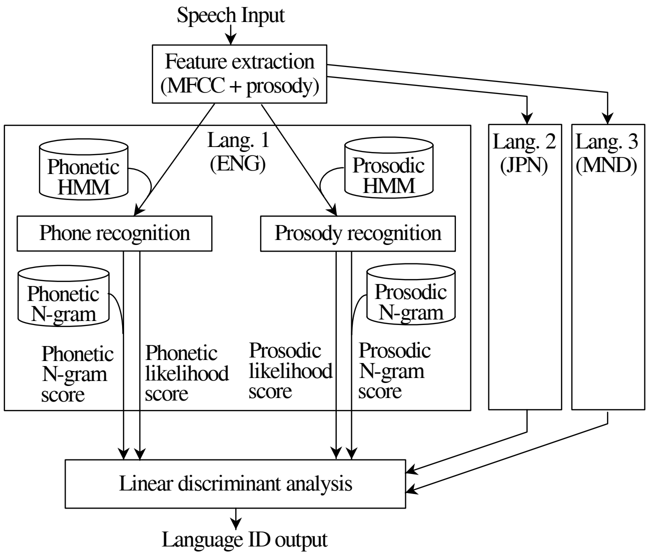

**Fig. 1** *System overview.*

When decisions are made for all pairs, a final decision must be made. The decisions are consistent in most three-language LID cases (one language beats the other two), but if there is any inconsistency between decisions, we apply an ad-hoc rule where the language that has the largest winning score becomes the final output.

## 4. FEATURE NORMALIZATION

We know that robustness in speech recognition systems can be achieved by applying some feature normalization techniques. Both cepstrum mean normalization (CMN) and mean and variance normalization (MVN) are popular. Recently, we proposed delta-cepstrum normalization (DCN) [8], which is a nonlinear transformation of cepstral coefficients using the cepstra and delta-cepstra histograms. It is known to be effective especially under noisy conditions. It is reasonable to expect that applying these algorithms to the phonetic features would improve the LID system performance. Furthermore, these algorithms may be applicable to prosodic features. If these algorithms can reduce irrelevant information such as personality and environmental fluctuations, LID performance would be improved.

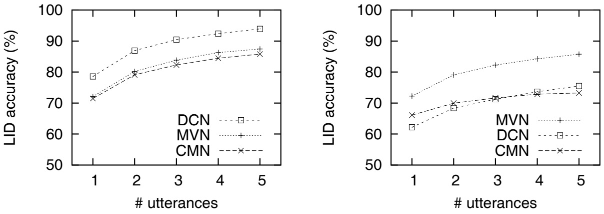

**Fig. 2** *LID accuracy of stand-alone systems. (left) Phonetic HMMs. (right) Prosodic HMMs.*

## 5. EXPERIMENTAL RESULTS

### 5.1. Databases and experimental setup

We created phonetic HMMs using phonetically labeled databases. We used the LDC Wall Street Journal database for English. For Japanese and Mandarin, we collected original databases that consisted of phonetically balanced sentences uttered by 120 (JPN) and 80 (MND) speakers. Prosodic HMMs were trained using the same databases without any labeling. The LDC Santa Barbara Corpus of Spoken American English and ASJ Japanese Newspaper Article Sentences databases were used as development/test data. We collected Mandarin development/test data that included free conversations between two people. For each language, 2692 utterances, with an average length of 3.0 seconds, were picked up. These utterances were divided into development and test sets, and we made experimental runs by switching their roles. The final LID accuracy was given by averaging the results. We also prepared noisy data by digitally adding the noise data taken from the JEIDA noise database to the test sets with SNR values of 0, 5, 10, 15, and 20 dB.

The phonetic HMMs consisted of 42 (ENG), 34 (JPN) or 57 (MND) phoneme models. Each model had three states, and each state had eight Gaussian mixtures. The prosodic HMMs consisted of 20 prosodic segment models. Each model had five states, and each state had one Gaussian distribution. We trained trigram language models for phonemes and prosodic segments by deleted interpolation, using the decoder outputs for the training data. The LDA parameters were calculated using the development data, in which each utterance contained only the language name label. Throughout all the experiments, the speech input was sampled by 16 kHz, and 13 MFCCs and three prosodic features were computed every 10 ms.

### 5.2. Stand-alone system experiments

The phonetic and prosodic HMM systems were first evaluated separately to study the contribution of each module to the complete system. The results are shown in Fig. 2. The LID performance for longer inputs was evaluated by adding the scores of successive utterances, weighed by the utterance length. Accordingly, the horizontal axis shows the number of utterances used for each test. The vertical axis represents the LID accuracy. Since the identification of three languages was forced, 1/3 is the theoretical lower limit of LID accuracy. We applied three feature normalization algorithms, CMN, MVN, and DCN, were applied to both the phonetic and prosodic features[1]. As the figure shows, applying DCN is helpful for phonetic HMMs, whereas it lowers the accuracy in prosodic HMM systems. The comparison of the best normalization algorithms revealed that the prosodic HMM system was not superior to the phonetic HMM system, but the LID accuracy can be as high as 85.8% when using 15 second speech (3 sec * 5 utt).

**Table 1**. *Average execution time for a one-second speech.*

| | | Phonetic HMM | Prosodic HMM |
|---|---|---|---|
| Feature extraction | | 0.022 s | |
| Recog. | ENG | 0.110 s | 0.015 s |
| | JPN | 0.086 s | 0.015 s |
| | MND | 0.154 s | 0.015 s |
| LDA | | less than 0.001 s | |

The average execution time was also measured using an Intel Pentium4 (2.66 GHz) processor running on the Linux operating system. The results are shown in Table 1. Since the number of models, the number of Gaussian mixtures, and the feature dimension were small, the prosodic HMM system was much faster than the phonetic HMM system.

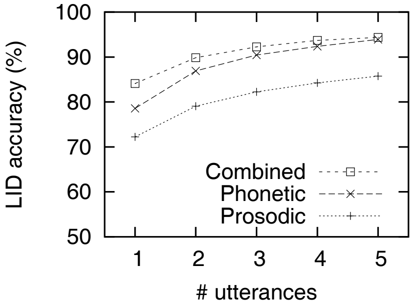

**Fig. 3** *LID accuracy comparison of phonetic, prosodic, and combined systems.*

### 5.3. Combined system experiments

Figure 3 shows the results obtained using the combined system. Taking the results of stand-alone experiments into consideration, we applied DCN to the phonetic features and MVN to the prosodic features. The combined system had an 84.1% LID accuracy for one utterance, which means a

[1] The terms CMN and DCN should not be normally used for prosodic features because they are not cepstra, but we used them for convenience of comparison.

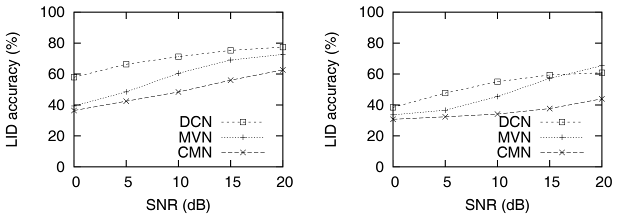

**Fig. 4** *LID accuracy of stand-alone systems under noisy conditions. Accuracies are for one utterance (3 seconds on average). (left) Phonetic HMMs. (right) Prosodic HMMs.*

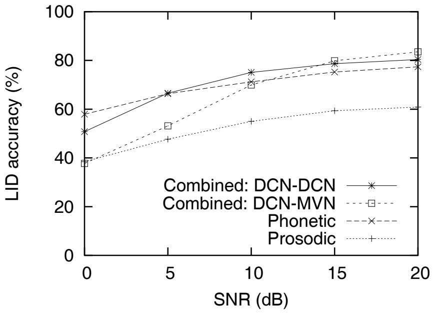

**Fig. 5** *LID accuracy comparison of phonetic, prosodic, and combined systems under noisy conditions.*

26% error reduction from the phonetic HMM system. As seen in Table 1, the execution time of the combined system for a one-second speech was 0.417 sec. This is only 12% longer than the time for the phonetic HMM system.

### 5.4. Noisy condition experiments

The proposed system was also evaluated under noisy conditions using artificially created data. In these experiments, we used the same LDA parameters calculated from the clean development data. Figure 4 shows the LID accuracy of the phonetic and prosodic systems. The horizontal axis represents the SNR, and all the results are for one utterance.

When applied to the phonetic features, DCN works effectively under noisy conditions, which is consistent with the previous speech recognition results. The benefits of MVN for the prosodic features decreases under low SNR conditions, while DCN improves. Nevertheless, the overall results suggest that the prosodic system is more vulnerable to noise.

Figure 5 shows the results of the combined system. Another series of experiments was carried out using the combined system in which DCN was applied to the phonetic and prosodic features. The LID accuracy of the DCN-MVN combined system, which was superior to the phonetic system under clean conditions, drops as the SNR decreases. However, the DCN-DCN combined system inherits the noise robustness of DCN applied to the prosodic features, and better results were obtained in SNR ranges between 5 to 20 dB.

## 6. CONCLUSIONS

In this paper, we proposed an LID system that uses prosodic HMMs. The system has an advantage because it requires no labeled corpus to create HMMs. The system is very fast, and its performance is quite reasonable, with an 85.8% accuracy in identifying three languages during approximately 15 seconds of speech. In addition, the combined system using phonetic and prosodic HMMs improves LID accuracy, where a 26% error reduction from phonetic HMM systems can be achieved with only 12% additional computation cost.

Feature normalization algorithms such as MVN and DCN have proven beneficial for LID. The phonetic HMM system was improved as in speech recognition systems. More interestingly, in some cases, normalization of the prosodic features also increased the LID accuracy.

Our experiments revealed that the prosodic HMM system is vulnerable to noisy conditions, even if DCN reduces the noise effect. Currently, F0 is estimated by a simple algorithm, so a robust F0 estimation algorithm must be introduced in the future to make our LID system more useful.

## 7. REFERENCES

[1] J. T. Foil, “Language identification using noisy speech,” *Proc. ICASSP*, Tokyo, Japan, 1986

[2] J.-L. Rouas, J. Farinas, and F. Pellegrino, “Automatic modelling of rhythm and intonation for language identification,” *Proc. International Congress of Phonetic Sciences*, Barcelona, Spain, 2003

[3] A. G. Adami and H. Hermansky, “Segmentation of speech for speaker and language recognition,” *Proc. EUROSPEECH*, Geneva, Switzerland, 2003

[4] M. A. Zissman, “Comparison of four approaches to automatic language identification of telephone speech,” *IEEE Trans. Speech and Audio Processing*, Vol.4, No.1, pp.31-44, 1996

[5] K. Iwano, T. Seki, and S. Furui, “Noise robust speech recognition using prosodic information,” *Proc. Workshop on DSP in Mobile and Vehicular Systems*, Nagoya, Japan, 2003

[6] T. Nagarajan and H. A. Murthy, “Language identification using parallel syllable-like unit recognition,” *Proc. ICASSP*, Montreal, Canada, 2004

[7] J. Gutiérrez, J.-L. Rouas, and R. André-Obrecht, “Fusing language identification systems using performance confidence indexes,” *Proc. ICASSP*, Montreal, Canada, 2004

[8] Y. Obuchi and R. M. Stern, “Normalization of time-derivative parameters using histogram equalization,” *Proc. EUROSPEECH*, Geneva, Switzerland, 2003