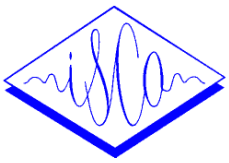

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

LANGUAGE IDENTIFICATION BASED ON SPEECH FUNDAMENTAL FREQUENCY

ITAHASHI Shuichi, DU Liang
e-mail: itahashi@milab.is.tsukuba.ac.jp
Institute of Information Sciences and Electronics
University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305
Japan

ABSTRACT

This paper describes a spoken language identification method based on speech fundamental frequency(F_0). The procedure is subdivided into three main stages: 1) F_0 extraction and segmentation; 2) polygonal line approximation of F_0 pattern; 3) discriminant analysis. The stage of F_0 extraction uses the Average Magnitude Difference Function(AMDF) and speech energy to estimate the fundamental frequency period of voiced speech sounds. In order to find better features from F_0 pattern, polygonal lines are used to approximate the F_0 contour of voiced intervals. After previous two stages, the complete parameter set is available for discrimination. The principal component analysis and discriminant analysis are performed at the last stage. The system is trained and tested using a CD-ROM "The Multi-language Speech Database for Telephonometry 1994", which is produced by NTT and NATC, and the OGI Multilanguage Telephone Speech Corpus.

Keywords: F_0 contour, Principal Component analysis, Discriminant analysis.

10.21437/Eurospeech.1995-357

INTRODUCTION

Many languages are used in the world. Most speech recognizers are designed to accept a single language today. However, it will be necessary for future speech recognizers to accept a variety of languages. Identification of spoken languages will play an important role as preprocessing for multi-language speech understanding [1-4].

Although it has long been recognized that prosodic information would contribute significantly to the language identification, there has been little methods so far using such information.

Characteristics of languages are represented by either segmental or prosodic features of speech. The fact that speech intonation plays an important role in spoken language understanding suggests that prosodic features could also be the basis of spoken language identification. In addition, fundamental frequency is supposed to be more robust than segmental parameters. In noisy environment like a subway train, we can hear accent or intonation of a conductor's announcement, even if it is difficult to recognize what he/she said. Since it is accepted that the variation pattern of F_0 is one of the best parameters to represent the prosodic features of spoken languages, here, we have tried to realize an automatic language identification system based on speech fundamental frequency[5-9].

The rest of the paper is organized as follows. In section 2, we briefly describe the speech material. In section 3, we present our analysis method: First is to extract the fundamental frequency (F_0) by the AMDF method, and to detect the voiced intervals using speech power. Second is to approximate the fundamental frequency contour with a set of polygonal lines for each voiced interval, and to minimize the mean square error between the lines and F_0 values; the optimum boundaries of the lines are determined using a dynamic programming procedure[10]. Third is to calculate statistical parameters from F_0 pattern and the polygonal lines. In section 4 we describe the discriminant analysis, section 5 is the analysis results and discussion, section 6 is conclusion.

SPEECH MATERIAL

The speech materials are composed of two groups:
Data Group One: this group is taken from a CD-ROM "The Multi-Lingual Speech Database for Telephonometry 1994"(MLSDT), which is produced by NTT and NATC, from which we selected 30-second utterances spoken by 4 different native male speakers in 6 different languages, which include Japanese, Chinese, Korean, English, French and German. In this case, the fundamental frequency contour of speech is extracted with a 16kHz the sampling frequency, 16 bit quantilization, 30 ms analysis window, and 10 ms frame interval.

MLSDT mainly aims at telephone call quality examination. It complies with the regulation of ITU-T recommendation P.80. It contains 21 languages. For each language, there are 8 native speakers, 4 males and 4 females, who are instructed to speak at ordinary speed and speech level.

Data Group Two: this group is taken from the OGI Multi-Language Telephone Speech Corpus(OGI-MLTS). In this case, we selected 20-second unconstrained spontaneous utterances produced by 5 different native male speakers in 6 different languages, which include Japanese, Chinese, Korean, English, French and German. And the experiment is developed under the conditions of 8kHz sampling frequency, 14 bit quantilization, 30 ms analysis window, and 10 ms frame interval.

The OGI-MLTS is designed to support research on automatic language identification and multi-language speech recognition. The utterances were spoken over commercial telephone lines by speakers of 11 languages. A total of 1545 calls, 246 in English,and an average of 122 calls in the remaining 9 languages, have been judged as useful after the evaluation. The ratio of male to female speakers was roughly 7:3 over all the languages. The corpus consists of up to nine separate responses from each caller, ranging from single words to short topic-specific descriptions to 60 seconds of unconstrained spontaneous speech[11].

ANALYSIS METHOD

Average Magnitude Difference Function (AMDF), which is represented as follows, is utilized for extracting the fundamental frequency(F_0).

$$D(m) = \frac{1}{N}\sum_{k=0}^{N-1}|x(k) - x(k+m)|$$

In consequence, the sum reaches a minimum in periodic signals with a delay $m = \Delta\tau$ which correspond to one F_0 period.

In order to detect voiced intervals, the short-time energy of the speech signal is used. For all frames, the intervals in which the speech energy is over a threshold are called voiced intervals. Actually, there are two thresholds, which correspond to the beginning and end of a voiced interval.

Speech intervals with speech powers greater than a certain threshold are detected automatically. Most F_0 extraction methods, including the AMDF method, sometimes make errors. Most of them are double-pitch or half-pitch errors. We have devised a simple error correction method. It calculates the mean F_0 for frames whose power is above a certain threshold. Raw F_0, $F_0 * 2$ and $F_0/2$ are compared with the mean F_0 and the one which is closest to the mean F_0 is adopted as a corrected F_0 value. Actually, $log_2 F_0$ is used in the following analysis:

The F_0 pattern is approximated by a set of polygonal lines which minimize the square error between each line and F_0 values. The optimum boundary of each line component of a polygonal line is determined by a dynamic programming procedure.

The approximation error decreases as the number of lines becomes large. It is necessary to approximate an F_0 contour with necessary and sufficient number of approximation lines to show typical characteristics of the F_0 contour of each language. One of the parameters we can use to determine the suitable number of component lines is the rate of decrease in the approximation errors. It does not necessarily give a globally optimal approximation. The number of lines should be determined according to the following conditions:

1. Keep the number of lines as small as possible.
2. Try to describe global characteristics rather than local features.

The suitable number of lines is determined so that the approximation error becomes smaller than a certain threshold.

The starting frequency, slope and duration of each component line are obtained, then their mean values, standard deviations (SD) and mean values of duration-weighted slope are calculated.

From the F_0 pattern, we extract 21 parameters for discriminant analysis:

1:	SD of F_0 of analyzed segments
2:	Skewness of F_0 of analyzed segments
3:	Kurtosis of F_0 of analyzed segments
4:	SD of P_0 of analyzed segments
5:	Skewness of P_0 of analyzed segments
6:	Kurtosis of P_0 of analyzed segments
7:	Correlation coefficient of F_0 and P_0
8, 9:	Number of Pl and Nl
10,11:	Mean slope of Pl and Nl
12,13:	SD of Pl and Nl
14,15:	Skewness of Pl and Nl
16,17:	Kurtosis of Pl and Nl
18,19:	Duration-weighted mean of Pl and Nl
20,21:	Relative starting frequency of Pl and Nl

Abbreviations utilized in the above table.

F_0:	Fundamental frequency
P_0:	Speech energy
Pl:	Approximation line of positive slope
Nl:	Approximation line of negative slope
SD:	Standard deviation

DISCRIMINANT ANALYSIS

Discriminant analysis is a statistical technique for classifying languages into mutually exclusive and exhaustive groups on the basis of a set of independent variables.

First of all, the extracted parameter set is used for principal component analysis to determine factors (i.e., principal components) in order to explain as much of the total variation in the data as possible with as few of these factors as possible.

It is customary to transform the raw data matrix to either a covariance matrix or a correlation matrix prior to principal component analysis. The primary

reason for our use of correlation matrix is that the extracted parameters have different units and scales.

Principal component analysis is given by:

$$Rx = \lambda x$$

where R denotes the correlation matrix; λ denotes the eignvalues.

Then, utilizing the results of principal component analysis, Mahalanobis distance discriminant analysis is performed to discriminate which group the input sample speech data belongs to.

Mahalanobis distance D_{ij}^2 is given by:

$$D_{ij}^2 = (x_i - x_j)' S^{-1} (x_i - x_j)$$

where x_i and x_j denote the respective vectors of measurements on stimuli i and j, and S is the pooled (within-group) variance-covariance matrix.

Closed and open experiments are developed at the stage of discriminant analysis. In the closed experiment, the same sample is used for training and discrimination. In the open experiment, the sample to be discriminated is excluded from training material. Both experiments of discriminant analysis have been quite well.

ANALYSIS RESULTS AND DISCUSSION

Table 1 shows the results of discriminant analysis.

Table 1: Results of discriminant analysis (%)

Experiment	*Data Groups One*	*Data Group Two*
Closed	89.1	96.7
Open	75.0	63.3

Figures 1 and 2 illustrate two examples of analyzed fundamental frequency values (dots) and approximated polygonal lines, which correspond to Data Group 1 and 2, respectively.

Figures 3 and 5 show the results of principal component analysis of 21 parameters on the first component(P1) and the second component(P2) plane, which correspond to Data Groups 1 and 2, respectively. Each point in the figures corresponds to each speaker.

Figures 4 and 6 indicate the results of principal component analysis of 21 parameters on the third component(P3) and the fourth component(P4) plane, which correspond to Data Groups 1 and 2, respectively. Each point in the figures corresponds to each speaker.

CONCLUSION

The results of the experiments proved effective, but there is room for the system improvement. Specifically, future work will include making experiments for female utterances, searching for better parameters to capture the features of F_0 contours, exploring the use of linear discrimination analysis for language identification.

ACKNOWLEDGEMENTS

This research is supported in part by Grant-in-Aid for Scientific research from the Ministry of Education Science and Culture No. 05241107 and in part by Real World Computing Partnership subcontract

References

[1] Y. Ueda and S. Nakagawa, "Prediction for Phoneme/Syllable/Word-Category and Identification of Language Using HMM", Proc. ICSLP90, Paper 27-6, pp.1209-1212 (1990).

[2] M. Sugiyama, "Automatic Language Recognition Using Acoustic Features", Proc. ICASSP91, Paper 16.S12.8, pp.813-816 (1991).

[3] Michael Savic, Elena Acosta and Sunil K. Gupta, "An Automatic Language Identification System", Proc. ICASSP91, Paper 16.S12.9, pp.817-820 (1991).

[4] Yeshwant K. Muthusamy and Ronald A. Cole, "Automatic Segmentation and Identification of Ten Languages Using Telephone Speech", Proc. ICSLP92, Paper pp.1007-1010 (Oct. 1992).

[5] S. Itahashi and T. Yamashita, "A discrimination method between Japanese dialects", Proc. ICSLP92, pp.1015-1018, Banff (Oct. 1992).

[6] K.Tanaka, S. Itahashi,"Discrimination Among Japanese Dialects Using Fundamental Frequency Pattern", (in Japanese), Tech. Rep. IEICE, SP92-114 (1992).

[7] S.Itahashi and K.Tanaka "A Method of Classification Among Japanese Dialects" Proc. Eurospeech93, Berlin, pp.639-642(1993).

[8] J. Zhou and S. Itahashi, "Feature Extraction for Spoken Language Discrimination Using Fundamental Frequency", Technical Rep. IEICE. Paper SP93-99, pp.61-66 (Nov. 1993).

[9] ITAHASHI Shuichi, ZHOU Jian Xiong and TANAKA Kimihito: "Spoken Language Discrimination Using Speech Fundamental Frequency", ICSLP 94, YOKOHAMA, pp1899-1902, 1994.

[10] S. Itahashi, "Description of Speech Data Patterns by Several Functions with Applications to Formant and Fundamental Frequency Trajectories", STL-QPSR 2-3/1978, pp.1-22 (1978).

[11] .K. Muthusamy, R.A. Cole and T.Oshika: "The OGI Multi-language Telephone Speech Corpus", Proceedings of the International Conference on Spoken Language Processing, Banff, Alberta, Canada, October, 1992.

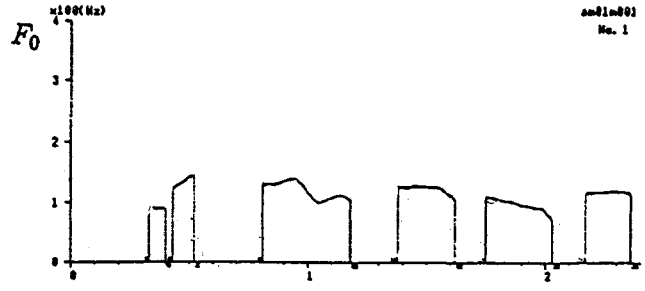

Fig1. F_0 contour and approximated lines.
(Data Group 1)

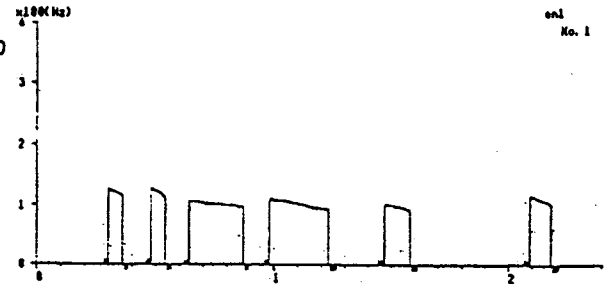

Fig2. F_0 contour and approximated lines.
(Data Group 2)

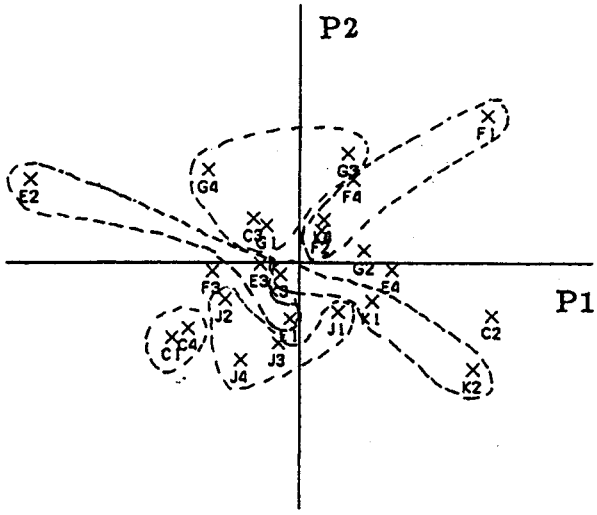

Fig3. Principal Component Analysis(P1-P2)
(Data Group 1)

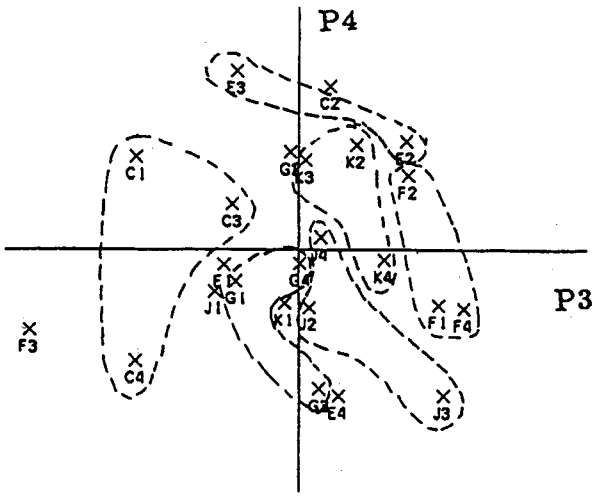

Fig4. Principal Component Analysis(P3-P4)
(Data Group 1)

Symbols	*Languages*
J1 - J5	Japanese
K1 - K5	Korean
C1 - C5	Chinese
E1 - E5	English
F1 - F5	French
G1 - G5	German

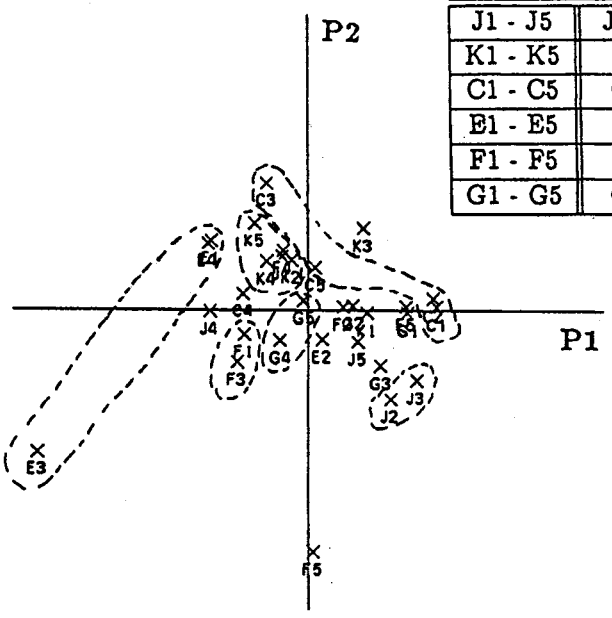

Fig5. Principal Component Analysis(P1-P2)
(Data Group 2)

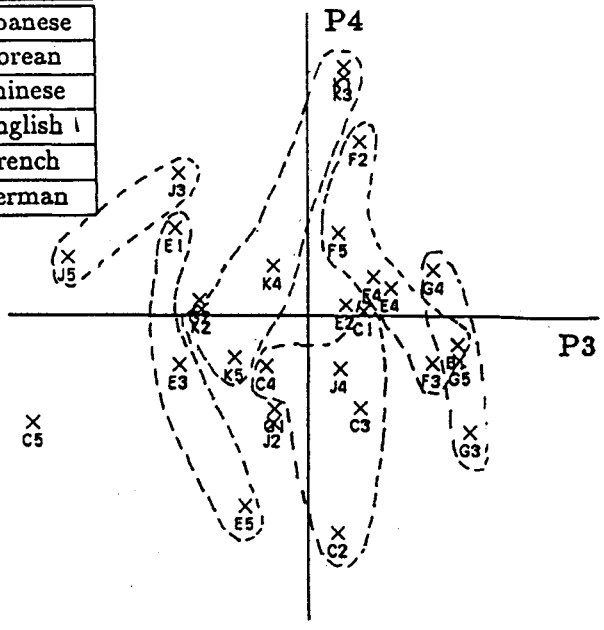

Fig6. Principal Component Analysis(P3-P4)
(Data Group 2)