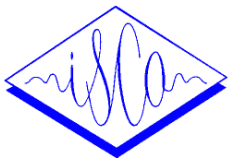

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A SPEAKER VERIFICATION SYSTEM USING PROSODIC FEATURES

B. Yegnanarayana*, S.P. Wagh** and S. Rajendran*

*Dept. of Computer Science and Engineering
**Dept. of Electrical Engineering
Indian Institute of Technology, Madras-600036, India

ABSTRACT

This paper describes a technique for automatic speaker verification based on prosodic knowledge in Hindi using neural networks. Properties of intonation patterns (changes in F0 as a function of time) and duration were exploited to extract speaker specific information from natural speech utterances, which were used for fixed text speaker verification task. A set of twenty-three features (fifteen pitch features and eight durational features) were extracted from a fixed natural utterance, using a word boundary hypothesization algorithm. A neural network model based on adaptive resonance theory (ART2) was used to verify speaker from the input feature set. The system was trained for twenty five speakers and tested with twenty seven impostors. The results show that the overall percentage of correct acceptance and correct rejection was found to be about 98%.

I. INTRODUCTION

In speaker verification the identity claim of a speaker is accepted or rejected by comparing a sample of his speech with the reference sample of utterance of the speaker, and making a decision on the basis of a predetermined similarity threshold. Speaker verification is an area where machine may surpass human performance. This is particularly true for unfamiliar speakers, where humans may take a longer time to learn a new voice, as compared to the time taken by a machine. Also the number of unfamiliar voices that can be retained in the short term memory of a human being is limited.

Several attempts were made in the area of speaker verification [1, 2, 3]. Most of the earlier work was based on linear prediction coefficients, reflection coefficients, logarithmic area coefficients and some dynamic features derived from speech waveforms. The classification techniques were based on various methods like hidden Markov model, vector quantization techniques, dynamic time warping, neural networks, etc.

In this paper we propose a method for speaker verification based on the prosodic features, pitch and duration in Hindi speech with an ART2 network for comparison and decision making. In the following Section we will discuss the design and development of the proposed system.

II. DEVELOPMENT OF THE SPEAKER VERIFICATION SYSTEM

Speaker verification generally consists of two stages: feature extraction and feature matching. Feature extraction includes data collection and data processing to get the desired features. Once the features are extracted, each speaker can be represented as a point in the feature space. In the matching stage the speaker is verified using the features from the sample utterance.

Fig.1 shows the block diagram for the speaker verification system. First the system has to be trained for all the customers (reference speakers) of the system. While testing the speaker verification system, the identity of claimed speaker is established by comparing the test sample with the reference samples of the speaker using a similarity threshold. The following sections describe the stages involved in the proposed speaker verification system.

1. Data Collection

We have collected speech data from twenty five adult speakers in text reading style. We have considered only cooperative speakers for data collection. Speech recording was done in an ordinary office environment in two different sessions with 2 to 3 weeks gap between the sessions. The test sentence selected for developing this system was */hama:re yahã: vivid^h b^{h}a:sa:ye boli: ja:ti: hai/* Twenty five utterances of the same sentence from each speaker were recorded. The speech was digitized using a 12 bit analog to digital converter at a sampling rate of 10 kHz. End points of the utterances were determined by a simple threshold logic on the amplitude of the speech signal or by visual inspection. Duration of the utterances are found to be varying from 2.2 to 3.5 sec. Pitch was extracted using the Simplified In-

verse Filter Tracking (SIFT) algorithm [4] and the gain contour was extracted using LP analysis [5].

2. Selection of Features

One of the most important steps towards achieving a successful speaker verification is the selection of features of speech which are capable of representing speaker dependent properties in the speech signal. The features selected should have high inter-speaker variability, low intra- speaker variability, resistance to attempted disguise or mimicry, robustness and measurability [6, 7, 8].

3. Features for the Proposed Speaker Verification System

The features are related to the fundamental frequency(F0) contour.They are the pitch frequency values at the valleys (initial syllable nuclei) and the peaks (final syllable nuclei) of the F0 contour for the monosyllabic, disyllabic and the trisyllabic words in the utterence of the test sentence [9]. In case of trisyllabic words, apart from the pitch frequency values at the valley and peak, the value of the pitch frequency at the middle syllable nucleus was also considered. For a monosyllabic word the value of the pitch frequency at the midpoint of the syllable nucleus was considered as a pitch feature. The durational features are the durations of the words and the total duration of the utterence of the sentence.

In the sentence */hama:re yahã: vivid^h b^{h}a:sa:ye: boli: ja:ti: hai/*, there are seven words, two are trisyllabic, four are disyllabic and one is monosyllabic. For the disyllabic words (*/yahã:/, /vivid^h/, /boli:/* and */ja:ti:/*) the pitch frequency values at the midpoint of the initial syllable (valley) and the final syllable (peak) were used as features while for the trisyllabic words (*/hama:re/* and */b^{h}a:sa:ye*) apart from the pitch frequency values at the valley and peak, the pitch frequency value at the midpoint of the middle syllable (*/-ma:-/* in */hama:re/* and*/- sa:-/* in */b^{h}a:sa:ye/*) were considered. In case of monosyllabic word (*/hai/*) the pitch value at the midpoint of the syllable was selected. Thus we have got fifteen pitch features, six pitch features corresponding to the two trisyllabic words, eight pitch features corresponding to the four disyllabic words and one feature corresponding to the monosyllabic word. The durations of of the seven words and the total duration of the sentence are expressed in terms of the number of analysis frames. Each frame is 25.6 ms with an overlap of 19.2 ms between successive frames. To derive the duration features, word boundaries were obtained using a word boundary hypothesization algorithm [9]. Thus a set of twenty-three features, fifteen pitch features and eight durational features were used for speaker verification.

Fig.2 illustrates the result of feature extraction for an utterance of the test sentence. The word boundaries obtained by the word boundary hypothesization algorithm are marked in the figure. The sentence has six boundaries and the algorithm has hypothesized all these boundaries correctly. Pitch contour is indicated by thick lines and the energy contour by a thin line. Vertical bars in the energy contour indicate the peaks in the energy contour. These peaks are used for determining the midpoints of the syllable nuclei. Dots marked in the pitch contour correspond to pitch frequency value at the midpoint of the syllable nuclei.

Robustness of the selected features

The main advantage of using the prosodic features for speaker verification is their robustness even under noisy input conditions. The robustness of the feature is mainly due to two reasons: First, the values of the pitch contour were used to derive the pitch accent features. Pitch contour is not affected by the spectral characteristics of the recording (such as level at which speaker talks, distance between microphone and speaker, etc.) and transmission systems. Hence it can be used more effectively in practical situations like recognition of speaker from telephone speech. Secondly, the word boundary hypothesization algorithm uses only gross parameters like energy and pitch frequencies in its implementation. Under noisy input conditions the correct computation of the entire pitch contour is difficult. But for hypothesization of the word boundaries only the high SNR portions of the speech signal need to be considered for extracting the pitch accent features. Under noisy input condition the pitch contour can be estimated using the properties of the group delay function [10]. The performance of algorithm is not affected by noisy input conditions because pitch accent pattern remains more or less the same. Hence the features are robust under noisy input conditions. In some cases errors occur due to errors in finding the peaks of the energy contour and due to errors in the pitch contour, which may result in the wrong placement of the boundaries. These errors can be corrected using the knowledge of the inherent durations of the words due to text dependent nature of the system.

III. NEURAL NETWORKS FOR SPEAKER VERIFICATION

Adaptive resonance theory (ART) architectures are neural networks that self organizes stable recognition codes in real time in response to arbitrary sequence of input patterns [11]. Development of this model was motivated by the function of brain which is able to receive new information as they arrive (plasticity) without changing the stability needed to ensure that the existing information is not erased or corrupted in the process. ART's continuous learning is one of its key benefits. Also, it does not require labeled training data and can

adapt to nonstationary data. The adaptive resonance theory paradigm(ART1) requires that the input vector be binary, whereas the ART2 paradigm is suitable for processing analog pattern. The proposed speaker verification system is based on ART2 network.

ART2 consists of an *attentional subsystem* and an *orienting subsystem.* The fields F1 and F2, as well as the bottom up and top-down adaptive filters, are contained within the ART's attentional subsystem. The two fields are: 1) *a feature presentation field* (F1) and 2) *a category presentation field (F2).*The pattern of activity that develops over the nodes of these layers F1 and F2 of the attentional subsystems undergoes cooperative and competitive interaction. These are called short term memory (STM) traces because they exist only in association with the application of an input vector. The weights associated with the bottom-up and top- down links between F1 and F2 are called long term memory (LTM) traces because the encoded information remains a part of the network for an extended period.

An auxiliary orienting subsystem becomes active when a bottom-up input to F1 fails to match the learned top-down expectation read out by the active category presentation at F2. In this case, the orienting subsystem is activated and causes rapid reset of the active category presentation at F2. This reset event automatically includes the attentional subsystem to proceed with the parallel search. Alternative categories are tested until either an adequate match is found or a new category is established. The search remains efficient because the search strategy is updated adaptively throughout the learning process.

1. Training and Testing of ART2 Network for Speaker Verification

The training set used to train the ART2 consists of input features without any target output. We have selected ten patterns for each speaker based on preliminary analysis. The clusters formed by the training with these patterns were classified into groups and each group corresponds to a speaker.There may be more than one group for a given speaker. During testing, the claimed speaker's identity will be verified using a similarity threshold. The speaker was identified from the unknown input pattern based on the cluster to which the pattern associates. If it fails to identify a proper match within the tolerance range determined by the predefined similarity threshold, the identity of the claimed speaker is rejected. An updating of clusters allocated to each speaker increases the verification accuracy significantly. This flexibility of the network helps us to add new speakers to the system without affecting the existing system.

Parameters of the orienting subsystem play a significant role in the performance of the system. The matching criteria is primarily determined by the vigilance parameter which controls the activation of the orienting subsystem. Higher vigilance imposes a stricter matching criterion and hence partitions the input set into finer categories provided all other parameters remain unchanged. Lower vigilance tolerates greater mismatches in the feature representation field and hence classification is done into coarser categories. Also, for every vigilance level the matching criteria was self scaling based on the complexity of the input patterns. For our system, we have decided to impose a higher vigilance(0.9) which gave sufficiently good clustering. But later we grouped these clusters to identify the speaker. The attentional gain control at F1 and F2 adjusts the overall sensitivity to pattern inputs and coordinates the different asynchronous functions of the ART subsystems.

IV. SUMMARY AND CONCLUSIONS

In this paper, a speaker verification system based on prosodic knowledge in Hindi using neural network was presented. The system was tested for twenty-five speakers. Twenty-five utterances of the sentence were collected from each speaker. All the speakers selected for data collection were native Hindi speakers. Ten utterances from each speaker were used for training of the system and the remaining fifteen utterances were used for testing. The speaker verification system was trained for twenty-five customers and tested with twenty- eight impostors. The overall percentage of correct acceptance was 98.

REFERENCES

[1] M. R. Sambur, "Speaker recognition using orthogonal linear prediction," IEEE Trans. ASSP, Vo. 24, pp.283-290, 1976.

[2] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Tras. ASSP, Vo. 29, pp. 254-272, 1981.

[3] G. R. Doddington, " Identifying speakers from their voices," Proc. IEEE, Vo. 73, pp.1651-1644, 1985.

[4] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," IEEE Trans. Audio and Electroacoustics, Vo. 20, pp. 367-377, 1972.

[5] J. Makhoul, "Linear prediction: A tutorial review," Proc. IEEE, Vo. 63, pp. 561-580, 1972.

[6] B. S. Atal, "Automatic speaker recognition based on pitch contours," J. Acoust. Soc. Amer., Vo. 52, pp. 1687-1689, 1976.

[7] F. Nolan, "The Phonetic Bases of Speaker Recognition," Cambridge University Press, NewYork, 1983.

[8] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," J. Acoust. Soc. Amer., Vo. 51. pp. 2044-2055, 1990.

[9] S. Rajendran and B. Yegnanarayana, "Word boundary hypothesization based on F0 patterns," To appear in Speech Communication.

[10] B. Yegnanaryana and V. R. Ramchandran, "Group delay processing of speech signals," Proc. ESCA Workshop on a Comparing Speech Signal Representation, pp. 411-418, Sheffild, England, 1992.

[11] G. A. Carpenter and G. Grossberg, "ART2: Self organization of stable category recognition codes for analog input patterns," Applied Optics, Vo. 26, pp.4919-4932, 1987.

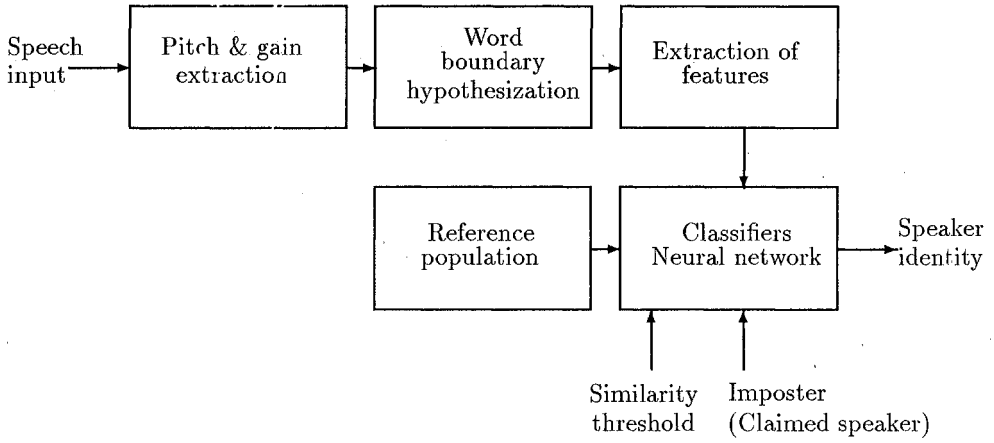

Fig.1. Block diagram of speaker verification system

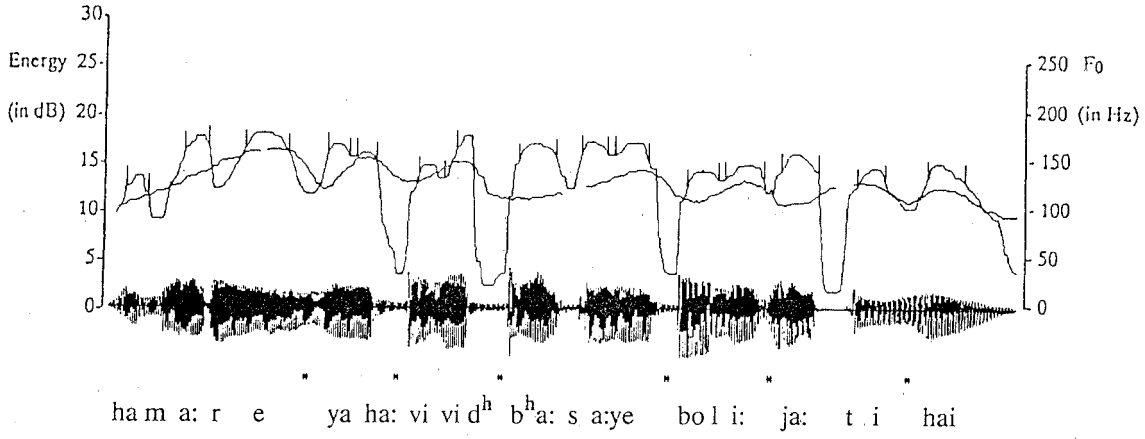

Fig.2. Hypothesization of word boundaries and feature extraction using word boundary hypothesization algorithm for the utterance, */hama:re yahã: vivid^h b^{h}a:sa:ye boli: ja:ti: hai/* 'Several languages are spoken here'. The pitch and gain contours are ploted over the waveform. Each pair of vertical bars on the gain contour indicates the location of syllabic nuclei. # indicates word boundary hypothesized by the algorithm.