Draft: Visual Pattern based Speech Recognition

PVS Rao and Sunil Kopparapu
Speech Group, Cognitive System Research Laboratory
Tata Infotech Limited, Mumbai

August 18, 2005

Abstract

This paper describes a novel spoken word recognition system based on visual pattern recognition. The main contribution of the paper is (i) converting a spoken word into a visual pattern, such that the pattern is invariant to change in intensity of the speech as well as any frequency shift and (ii) the comparision of visual patterns which is equivalent to speech signal comparision.

1 Speech as Visual Pattern

A speech signal can be represented as a spectrogram. A spectrogram is a 3D plot as a 2D image, the x-axis represents the time component, while the y-axis represents the frequency component and the intensity of the frequency component at any given time instant is represents on z-axis (intensity of the image). This spectrogram is usually discussed and used in speech signal processing in literature. We would like to use the spectrogram as the visual pattern to aid speech recognition.

The visual pattern should be invariant to any change in the intensity of the spoken speech and in addition should be invaiant to any frequency multiplication that might happen in the speech signal. To take care of these two invariance requirement, we plot the spectrogram in the log scale along the frequency axis. This representation enables any frequency multiplication that might occur in a frequency domain of speech signal into translation in the log-frequency domain - making the representation invariant to frequency multiplication. The log-frequency spectrogram can be represented as a contour plot. The contours could be hill climbing or hill descend in the log-frequency spectrogram. The effect of visualizing the spectrogram as a contour plot is that any amplitude scaling of the speech signal (intensity variation) would not effect the contour map significantly except may be in the form of an additional contour be appended to the contour map or a contour being removed from the contour map. In this sense the contour representation of the speech signal is invariant to the change in intensity of the speech signal. So any speech signal can be represented as a log-frequency contour spectrogram (LFC-Spectrogram).

2 LFC-Spectrogram Measure

Define a measure (essentially a cost) to convert/transform the LFC-Spectrogram of a signal $x(n)$ to the LFC-Spectrogram of another signal $y(n)$. This cost would determine the closeness of signal $x(n)$ to signal $y(n)$. The exact cost measure needs to be worked out in detail.

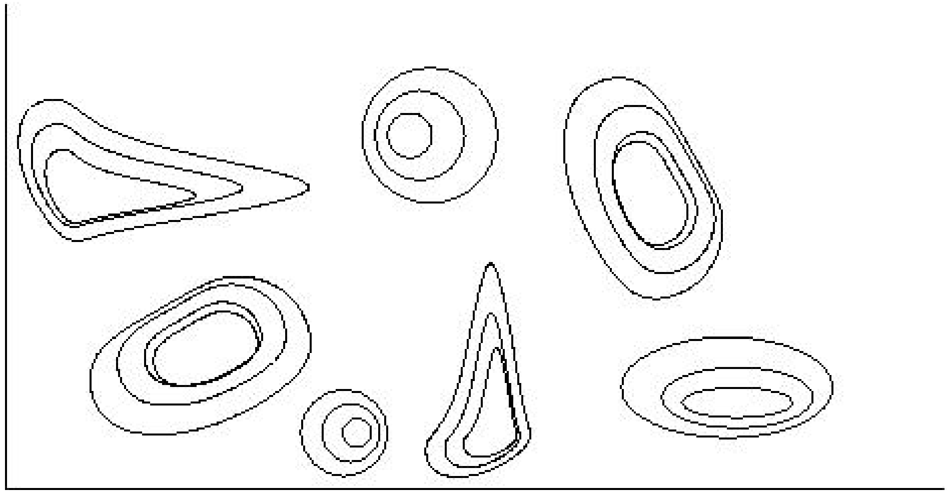

Figure 1: Example of a log-frequency contour spectrogram. The x-axis is the time and the y-axis is the log frequency. The contours depict the intensity of the energy at that frequency and that time.

3 Speech Recognition

Each speech sample would be represented as its LFC-Spectrogram. Given the properties of LFC-Spectrogram speech signals representing the same word would have similar contour map compared to the LFC-Spectrogram of a speech signal corresponding to another spoken word. This is the basis for using LFC-Spectrogram for speech recognition.

Given a set of N words that need to be recognized by the system, obtain k representative speech samples of the N words. Calculate the LFC-Spectrogram of all the k samples of each of the N different words. Now given the speech sample of a spoken word, sat T (test sample). We need to determine to which of the N words the word T is closest to. First, we find the LFC-Spectrogram of the signal T. Then we determine the cost of converting the LFC-Spectrogram of T to LFC-Spectrogram of word n, call it C_{Tn}. We determine

$$C_{Ti} \quad \text{for} \quad i = 1, 2, \cdots, N$$

and find l such that

$$C_{Tl} < C_{Ti} \quad \text{for} 1 \leq i \leq N \quad \text{and} \quad i \neq l$$

4 Conclusions

New approach which uses visual signal processing

5 Things to do?

- Crucial step - developing a metric that cab reliably compare two contours maps. Basically identify the penalty to convert one contour into another

- There might be some measure in literature of image processing. Check **contour matching**, **water sheding**, **Earth movers distance** literature. These might not be usable directly though.

Location: 163.122.72.227: /home/nil/projects/speech/doc/visual_sr
Document Last updated on La updated 20/09/2005 (10:19) by Sunil