

DETECTION OF HUMAN SPEECH IN STRUCTURED NOISE

John D. Hoyt
Engineering Research Facility
Federal Bureau of Investigation
Quantico, VA 22135

Harry Wechsler
Dept. of Computer Science
George Mason University
Fairfax, VA 22030

ABSTRACT

This paper describes research to develop an efficient system that provides a binary decision as to the presence of speech in a short (one to three second) time sample of an acoustic signal. A method which is efficient and reliably detects human speech in the presence of structured noise (such as wind, music, traffic sounds, etc.) is described.

Two separate algorithms were developed. The first algorithm detects the presence of speech by testing for concave and / or convex formant shapes. The second algorithm is a statistical pattern classifier utilizing radial basis function (RBF) networks with mel-cepstra feature vectors. Classification errors are not consistent across these two different methods. As a consequence, we plan to reduce our error rate by fusion of these methods.

1. INTRODUCTION

The problem area of interest in this research is in providing a binary yes / no decision that human speech is present in a time segment of an acoustic signal of one to three seconds in duration. The desire is for a method which is independent of both the language and the speaker, and (ultimately) can be inexpensively implemented in battery powered equipment. For this research, noise is defined as any sound that is not speech (for simplicity, we ignore singing as either speech or non-speech). This collections of sounds we call structured noise, as it includes quasi-periodic signals, such as music; pure tones; impulse noise; and Gaussian like sounds, such as running water.

In order to freely interact with a computing environment, to the extent that the computer is truly embedded within the environment, it is desirable to communicate with the computing resources by voice and gestures, without utilizing encumbering close talking microphones, keyboards, mice, etc. Research into microphone array systems has demonstrated the ability to obtain a high quality speech signal in a room environment [1]. However, in order to track the speaker in anything but

a quiet room, it is necessary to be able to determine which signals are human speech.

If multimedia personal communications devices (PCD's) are to become truly accepted by the general public, the PCD must be small, unobtrusive, and easily used. Simple commands and queries via speech rather than a keyboard would be preferable. However, if the PCD is to be small and unobtrusive then it would be best if it were battery powered and did not require the user to wear a close talking microphone. These two requirements imply that the PCD must be energy efficient and be capable of dealing with acoustic signals that are not human speech. A low power method of detecting the presence of speech could conceivably be utilized to initiate more computationally intensive (and thus more power intensive) speech to text recognition algorithms.

Much of the pervious work in the area of speech detection falls into two categories. The first is that of a speech signal in a quiet environment, where a simple energy measure is sufficient to classify speech as being present. The second is in determining that speech is present on a communications channel. In this case, the speech signal has statistical properties that differ greatly from static or data communication signals. Existing speech detection algorithms designed for a communications environment utilize various combinations of energy estimation, zero crossings, correlation functions, LPC coefficient and power change ratios, or bandpass filterbank signal to noise ratio thresholds [2][12][13]. These methods will not work when the noise has a long term power spectral density similar to speech and periodicity similar to speech (such as instrumental music).

2. DATA AND FEATURES

The source of speech used for training and testing is the TIMIT acoustic - phonetic continuous speech corpus [3] developed by the Advanced Research Projects Agency (ARPA). This corpus consists of 630 different speakers which are divided approximately evenly between male and female, and between the eight major dialects of American

English. Noise sources (rain, traffic sounds, rifle fire, bird songs, surf, etc.) are from a two volume compact disk (sound effects) [4] and from both compact disks and recordings from the reception of commercial FM radio stations (instrumental music - classical, jazz, and rock).

By examining the data, it was observed that all speech signals provided convex and concave formant shapes in frequency bands between 4 KHz and 400 Hz. The only non-speech signal evaluated that produced formant shapes that met the same criteria was a police siren. Instrumental music produced autocorrelation results and power spectral estimates that were indistinguishable from speech. These observation led to the formant tracking algorithm.

3. FORMANT TRACKING ALGORITHM

The concave and convex format patterns were observed in speech, but not noise samples. Thus it was decided to construct a system which performs an analysis of the shape of formants. In order to smooth the spectrograms to allow a simple peak detection algorithm to locate formants, a 20 pole linear predictive coding (LPC-20) [5][11] file was made from the time series data, and the LPC file was then used as input to a discrete Fourier transform (DFT) [6]. Peaks in the resulting smoothed time - frequency data are the features utilized by this algorithm. Figure 1 is a block diagram of this algorithm.

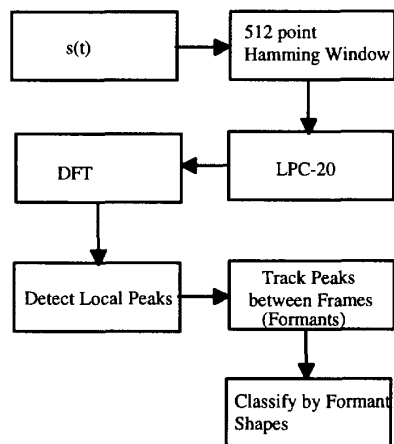


Figure 1: Formant Tracking Algorithm

The algorithm searches the peak data from the smoothed spectrograms for convex, concave and straight sections utilizing a type of chain code algorithm [14]. These shapes must meet a minimum length (duration) criteria, or they are discarded as noise. The minimum

length of each shape section (concave, convex, or straight) was experimentally varied to maximize the detection of speech. Other parameters that must be adjusted include; a maximum length for each shape section in a formant (to prevent false positives for such sounds as a police siren); a maximum difference between peaks in adjacent frames that will be included in the same formant; and a minimum length (time duration) for a formant. The surprising part of our research is not that the concave / convex formant shapes represent speech. The psycho-acoustic literature shows that this has been known for some time [7]. What is interesting is that these shapes do not appear reliably in non-speech sounds, including music. Structured noise produces either no formants, or formants which are straight (but not necessarily level).

4. RADIAL BASIS FUNCTION ALGORITHM

A block diagram of this algorithm is shown below in figure 2. This algorithm was developed utilizing purely statistical methods with unsupervised clustering and supervised training data. The preprocessing stage of this algorithm is the formation of mel-cepstra [5] [8] feature vectors from the basic time series acoustic data. Mel-cepstra were chosen as they are the feature vectors utilized in a large number of current speech recognition systems. Thus, by utilizing the same preprocessing, an additional processing burden would not be incurred in a device such as a PCD if speech were detected, and speech recognition was desired.

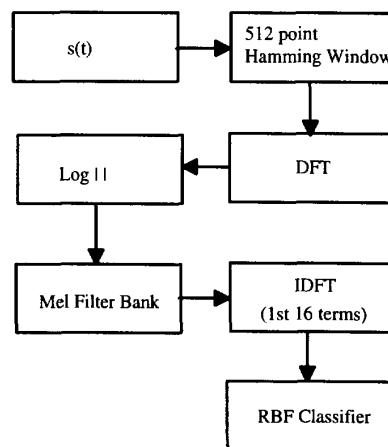


Figure 2: Mel Cepstra / RBF Algorithm

As the distributions of the speech and non-speech acoustic signals are, in general, unknown, a nonparametric classifier is required. Previous work by Ng and Lippmann

had shown that Radial Basis Function (RBF) classifiers performed well in classification problems of high dimensionality and RBF classifiers train quickly, classify quickly, and require modest amounts of computational resources [9][10]. For a RBF classifier, the input feature vectors are projected by the basis functions onto a high dimensional space. In our classifier, these basis functions are hyper-elliptic Gaussian functions. A linear combination of these Gaussians defines a hyperplane that divides the output into the various classes [9][10]. K-means clustering was utilized to determine the RBF cluster means. The variance of each cluster is set equal to the variance of the Euclidean distances of all of the class members within a cluster from the cluster mean. Additionally, the variance is calculated separately for each dimension in the feature space. Our mel-cepstra feature vectors are generated such that each vector represents 16 ms of the time series data. As we are interested in detecting speech over roughly a second interval, we store the most recent 32 outputs of our classifier, and declare the signal as speech or non-speech based on the number of these 32 values that exceed a threshold value.

5. TRAINING AND TESTING

The RBF algorithm was trained on one second speech samples taken from twenty (10 male and 10 female) speakers chosen at random from the TIMIT training data, as well as 25 non-speech samples chosen at random from the CBS [4] compact disk set and commercial FM radio broadcasts. Our formant tracking algorithm was tuned to maximize its performance over the same training data. All data was digitized at 16K samples per second with 16 bits per sample. All testing was performed on data that was not utilized in training. All data was preprocessed to remove periods of silence.

In the table below, noisy speech data was formed by adding TIMIT speech with digitized non-speech sounds. Six female samples and six male speech samples were taken at random from the TIMIT test data. The noise (non-speech) recordings were digitized from the CBS [4] compact disk set and from music digitized from commercial FM radio broadcasts and compact disks. In all cases, the test speech and noise samples were not utilized in training. A 1964 song, *Leader of the Pack*, was also digitized for speech / non-speech testing. Vocal sections of this song are noted as LOP_V and non-vocal as LOP_M in the table below. Results in the table are given in percent of correct classification as either speech or non-speech. Music in the speech + music sections of the table below were four different instrumental sections - two classical and two jazz.

Data	Formant Algo.	RBF Algorithm
clean speech	83.7%	75.0%
classical music	100%	100%
jazz / rock	66.7%	100%
cabin of 747 in flight	100%	100%
police siren	100%	100%
White Noise	100%	100%
heavy rain	100%	100%
LOP_V	100%	100%
LOP_M	100%	100%
speech + music 4:1 SNR	75.0%	66.7%
speech + music 2:1 SNR	33.3%	66.7%
speech + music 1:1 SNR	33.3%	66.7%
speech + rain 4:1 SNR	66.7%	66.7%
speech + rain 2:1 SNR	0%	0%
speech + WN 4:1 SNR	0%	91.7%
speech + WN 2:1 SNR	0%	41.7%

5. CONCLUSIONS

Our two algorithms reliably distinguish between speech and environmental sounds. When speech is mixed with environmental sounds (music and the sound of heavy rain were tested), detection remains comparable to clean speech until the SNR is less than approximately 4:1. As the two algorithms mis-classify different data, fusion of the two is the next step for our research.

White noise was also added to our test speech samples. Results were not indicative of the performance using additive environmental noise. We believe that this is due to the $1/f$ nature of spectral power estimates for the environmental noises utilized.

6. REFERENCES

- [1] J. L. Flanagan and H. F. Silverman, Co-Chair, "Material for International Workshop on Microphone Array Systems", Technical Report LEMS-113, Brown University, October 1992.
- [2] J. Burgett, "Description of the Readability Algorithm," UNCLASSIFIED Working Paper, Department of Defense, September 1986.
- [3] *TIMIT Acoustic - Phonetic Continuous Speech Corpus*, NIST Speech Disk 1-1.1, NITS Order No. PB91-505065, October 1990.
- [4] *The New CBS Sound Effects Library*, CBS Special Products, 51 W 52 St., NY, NY, 1977.
- [5] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice - Hall, 1978.
- [6] A. V. Oppenheim and R. W. Schafer, *Discrete - Time Signal Processing*, Prentice - Hall, 1989.
- [7] B. Moore, *An Introduction to the Psychology of Hearing - 3rd Ed.*, Academic Press, 1989.
- [8] D. O'Shaughnessy, *Speech Communication, Human and Machine*, Addison - Wesley, 1987.
- [9] K. Ng and R. Lippmann, *A Comparative Study of the Practical Characteristics of Neural Networks and Conventional Pattern Classifiers*, Masters Thesis, MIT, May 1990.
- [10] D. Hush and B. Horne, "Progress in Supervised Neural Networks, What's New Since Lippmann?", *IEEE Signal Processing Magazine*, January 1993.
- [11] J. Dieller, J. Proakis, and J. Hansen, *Discrete - Time Processing of Speech Signals*, Macmillan, 1993.
- [12] W. Chan and D. Faconer, "Speech Detection for a Voice / Data Mobile Radio Terminal," *Proc. of IEEE Conf. on Comm.*, pp F8.7.1-F8.7.5, June 1983.
- [13] J. Yang, "Frequency Domain Noise Suppression Approaches in Mobile Telephone Systems," *Proc. of IEEE ICASSP 1993, Vol II*, pp 363-366, April 1993.
- [14] R. Gonzalez and P. Wintz, *Digital Image Processing*, Addison-Wesely, 1987.