REAL-TIME DISCRIMINATION OF BROADCAST SPEECH/MUSIC

John Saunders

Sanders, A Lockheed Martin Co.
Nashua, NH 03061

ABSTRACT

We describe a technique which is successful at discriminating speech from music on broadcast FM radio. The computational simplicity of the approach could lend itself to wide application including the ability to automatically change channels when commercials appear. The algorithm provides the capability to robustly distinguish the two classes and runs easily in real time. Experimental results to date show performance approaching 98 % correct classification.

1. MOTIVATION

Many listeners are uninterested in the commercial and talk programming on broadcast radio, prefering music instead. Often they are unable to "surf" channels as a period of music segues into a segment of talk and commercials. A reliable, low cost, automatic program monitoring and discrimination function built in to a radio set would be a desirable feature. It could conceivably be integrated with the station scan control function to skip to the next station within a pre-defined set if the content does not match the user's choice.

Other applications include the long-term monitoring of a given station(s) to determine how much airplay is dedicated to music, possibly to ensure compliance with broadcast requirements. Surveillance receivers could employ this capability to monitor several radio channels, more effectively ferreting out intelligence by ignoring channels playing music.

We seek a simple technique for discriminating music from speech that is general in nature so that it applies across all the multitudinous forms of music yet requires minimal computation compared to other methods [1]. First, some of the salient differences of speech and music must be examined. A few of these features [2] might be:

- **Tonality.** Music tends to be composed of a multiplicity of tones, each with a unique distribution of harmonics. This pattern is consistent regardless of the types of music or instruments. Speech exhibits an alternating sequence of tonal and noise-like segments.

- **Bandwidth.** Speech is usually limited in frequency to about 8 KHz whereas music can extend through the upper limits of the ear's response at 20 KHz. In general, most of the signal power in music waveforms is concentrated at lower frequencies.

- **Excitation patterns.** The excitation signals (pitch) for speech usually exist only over a span of three octaves while the fundamental music tones can span up to six octaves.

- **Tonal duration.** The durations of vowels in speech is very regular, following the syllabic rate. Music exhibits a wider variation in tone lengths, not being constrained by the process of articulation. Hence, tonal duration would likely be a good discriminator.

- **Energy sequences.** A reasonable generalization is that speech follows a pattern of high energy conditions of voicing followed by low energy conditions which the envelope of music is less likely to exhibit.

2. ALGORITHM DESCRIPTION

In general speech patterns tend to contain a very regular structure. Words are made up of syllables and syllables are made up of consonant clusters followed by vowels and then likely followed by trailing consonants. It is easy to visually discern these patterns from the acoustic waveform. Vowels are usually high energy events owing to the lack of vocal tract constriction and show a visible periodic component due to the glottal excitation. Most of the spectral energy is contained at low frequencies. Consonants are likewise distinguishable, being produced by frication, a constriction in the vocal tract causing a turbulent air flow, attenuation

0-7803-3192-3/96 $5.00©1996 IEEE

of acoustic energy and damped resonances. The resulting speech is noise-like, producing spectral energy distributed more towards the higher frequencies.

The characteristic structure of speech is a succession of syllables composed of short periods of frication followed by longer periods of vowels or highly voiced speech. This simple but widely accepted generalization can be used to positively recognize a waveform as being predominantly speech.

Several features can be developed to help recognize this pattern. One of the most indicative and robust measures to discern voiced speech is the average zero-crossing rate (ZCR) of the time domain waveform. It provides a measure of the weighted average of the spectral energy distribution in the waveform, the spectral center of mass. The dominant frequency principle [3] shows that when a certain frequency band carries more power than other bands, it attracts the normalized expected number of zero crossing per unit time. As a result, the ZCR has been widely used in practice as a strong measure to discern fricativity from voiced speech.

The effectiveness of the ZCR and the attendant dominant frequency principle to the problem of discriminating speech from music can be understood. While both music and commercials are mainly singing with musical accompaniment, we observe that commercials strongly emphasize the speech component, a condition less likely to occur for music. With a mixture of speech and music, the signal energy contribution of speech is greater than the contribution of the music in order to emphasize the message of the talker. Conversely, it is observed that for music the singing content does not typically tend to dominate the music content. In a general sense, commercial and talk radio programming are strongly speech-like while music programming is not. The ZCR per unit time is effective at following the dominant component of the spectral composition and hence can be a good discriminator. The signal suppression effect [4] of the non-linear limiting operation inherent in zero-crossings measurement plays a role in emphasizing the stronger component of the waveform.

Examples of the ZCR contour for speech and music are shown in Figures 1 and 2, respectively.

Speech signals produce a marked rise in the ZCR during periods of fricativity occuring at the beginning and end of words. Music does not show such abrupt increases of the ZCR, it being largely tonal. Its ZCR variation is not distributed with the same bimodality as voiced and unvoiced speech. This bimodality skews the ZCR distribution towards the high end, causing several points in the distribution to lie at values significantly higher than the mean. These events are due to strongly

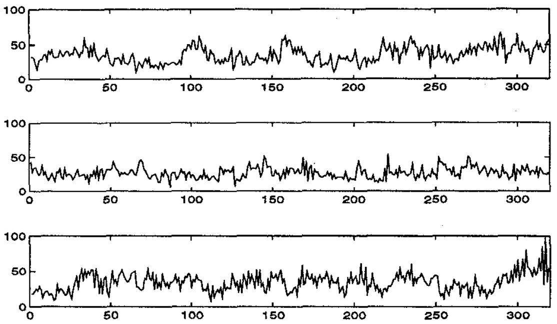


Figure 1: Zero crossing rate contour for music

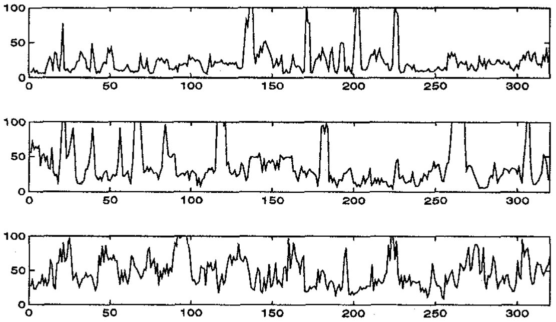


Figure 2: Zero crossing rate contour for speech

unvoiced speech (unvoiced fricatives, affricates). An algorithm that can capture this bimodality would be successful as a music/speech discriminator.

The energy contour of a waveform is well known to be capable of separating speech from music. The contour of music tends to show a much smaller number of dips and peaks than speech and it quite often shows little change over a period of several seconds. The alternation between voicing and frication in speech produces a marked change in its energy contour.

The algorithm is primarily based on the lopsidedness of the distribution of the ZCR. The first step is to measure the ZCR of the signal over a 2.4 second segment of the data. Next, obtain the mean ZCR and define a fixed threshold a certain number of units both above and below the mean value. The last step is to count the difference between the number of points at the low end of the distribution below the lower limit and the number of samples that exceed the higher limit. If this statistic exceeds a specific threshold, the distribution outside these bounds is significantly skewed and the waveform is likely speech.

3. EXPERIMENTS

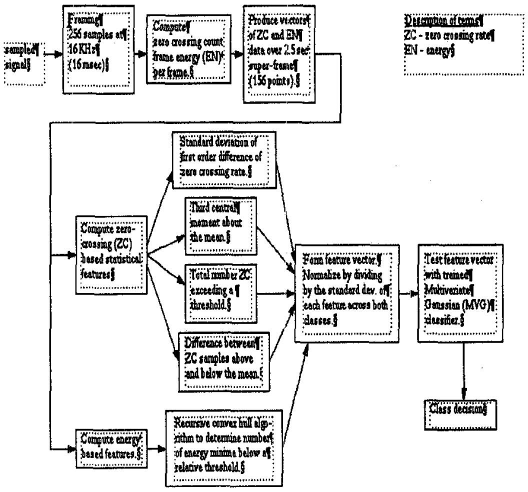
A sample rate of 16 KHz was chosen for this discrimination technique. Non-overlapped frames of data consisting of 256 samples at 16 millisecond intervals were formed. The total number of zero-crossings within this interval were counted and the energy was collected by summing squared samples. A block of 2.4 seconds containing 150 frames was then used to compute statistically-based features. A multivariate-Gaussian classifier [5] separates feature space and decides the class of the test token.

Experiments were performed with measures of the skewness of the distribution of the ZCR. These features were the standard deviation of the first order difference of the zero crossing rate, the third central moment about the mean, the total number of zero crossings exceeding a threshold, and the difference between the number of zero crossing samples above and below the mean. An important step was then to normalize the features by dividing by the standard deviation of each feature across both classes. We obtained classification performance averaging 90 % using these features.

Improved performance was obtained by including an energy contour dip measure into the discrimination process. A recursive convex hull algorithm used for syllabic segmentation [6] was employed to determine the number of energy minima below a given threshold which was relative to the peak energy within the collection interval. By adding this feature as a another dimension to the classifier, as shown in Figure 3, performance improved to an average of 98%. The potential class separability is revealed by testing on the training data. This test resulted in 98.4% class separation.

Data was collected manually by listening, collecting and storing features, and labeling the segment. A variety of content was processed, including talk, commercials, and many types of music. Once the classifier was trained, the parameters were stored and fed into the real-time feature extraction/classifier routine. The experimental setup used a Gradient A/D [7] unit attached to a workstation. The real-time data captured by the Gradient unit was transfered to the workstation which graphically indicates the class decision for each collection interval. Using data processed on the fly and tuning the radio dial at will, the classification accuracy averaged between 95% and 96%.

As an example application of this technique, station WZLX-FM, 100.7 MHz in Boston was monitored between 12 PM and 2 PM on April 8, 1995. The graph in Figure 4 shows that approximately 71.8 % of the airplay is dedicated to music and the remainder to talk and commercials.



sampled¶ signal§
Framing¶ 256 samples at¶ 16 KHz¶ (16 msec)§
Compute¶ zero crossing count frame energy (EN) per frame.§
Produce vectors¶ of ZC and EN¶ data over 2.5 sec super-frame¶ (156 points).§
Description of terms¶ ZC - zero crossing rate¶ EN - energy§
Standard deviation of first order difference of zero crossing rate.§
Third central¶ moment about the mean.§
Compute zero-crossing (ZC) based statistical features§
Total number ZC exceeding a ¶ threshold.§
Difference between¶ ZC samples above and below the mean.§
Form feature vector.¶ Normalize by dividing by the standard dev. of¶ each feature across both classes.§
Test feature vector with trained¶ Multivariate¶ Gaussian (MVG)¶ classifier.§
Class decision§
Compute energy based features.§
Recursive convex hull algorithm to determine number¶ of energy minima below a¶ relative threshold.§

Figure 3: Block diagram of processing flow.

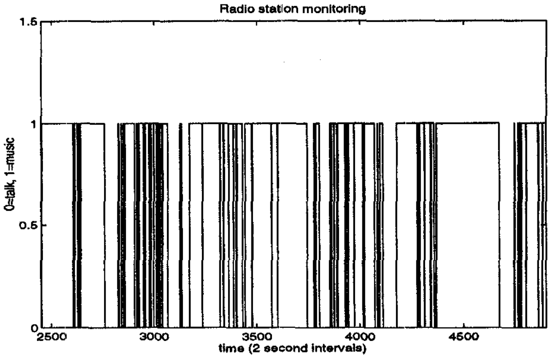


Figure 4: Talk/speech monitoring of a radio station

4. CONCLUSION

The minimal computation associated with this type of program content decision processing lends itself to implementation in radio receivers at very low add-on cost. The strict use of time-domain features avoids the need for an FFT processor to compute spectral features. Given that this algorithm has only been developed in an exploratory fashion, it is expected that improved performance beyond that reported is possible.

5. REFERENCES

[1] Hoyt J., Wechsler, H. "Detection of Human Speech in Structured Noise", Proc. ICASSP-94, Vol. II, pp.237-240.

[2] Backus, J., *The Acoustical Foundations of Music*, Murray, London, 1970.

[3] Kedem, B. "Spectral Analysis and Discrimination by Zero-Crossings", Proceedings of the IEEE, Vol. 74, No. 11, November 1986.

[4] Gardner, F.M., *Phaselock Techniques*, John Wiley & Sons, Inc., New York, NY, 1979.

[5] Tou, J.T., and Gonzales, R.C., *Pattern Recognition Principles*, Addison Wesley, 1974, Reading, MA.

[6] Mermelstein, P., "Automatic segmentation of speech into syllabic units", Journal of the Acoustical Society of America, vol. 58, No. 4, Oct. 1975.

[7] *DeskLab*^{TM}216 User Manual, Gradient Technology, Inc. , Burlington, NJ, 1992.