SILENCE AS A CUE TO RHYTHM IN THE ANALYSIS OF SPEECH AND SONG

David Gerhard
Computer Science, University of Regina, 3737 Wascana Pkwy, Regina, SK Canada, S4S 0A2. david.gerhard@uregina.ca

1. INTRODUCTION

The *rhythm* of a piece of sound is a measure of the existence, duration and repetition of time events in the sound. When dealing with full spectrum instrumental music, these time events are readily identified using spectral analysis techniques such as filter banks. Rhythmic structure is often apparent in percussion instrumentation, for example drums and cymbals, which are normally in different spectral locations than the rest of the instrumentation.

When dealing with human utterances, the spectral filter bank methods are not as successful because the acoustic events that may contribute to a sense of rhythm, are similar in spectral content to the rest of the utterance. Different methods must be employed to identify rhythmic structure. This paper discusses the merit of using silence as a cue to the rhythmic structure of human utterances, specifically speech and song. Identifying the rhythm of a human utterance applies to the problem of music information retrieval, specifically in the instance where a human musical utterance (i.e. song) is used as a query in a musical database. Currently, a string of note directions is commonly used as a search string into the database, without referring to rhythmic information. Rhythmic cues may help in retrieval accuracy. Further, rhythmic information may aid in he differentiation between speech and song, applicable to speech recognition systems as well as speech and music therapy.

This work seeks to characterize the statistical distribution of sound pressure level (*spl*) across an utterance, identifying silence as a value of *spl* below a hysteresis threshold. This work concentrates on English speech and the western musical idiom. Several researchers have identified a 4 Hz power modulation in speech while attempting to differentiate between speech and music [2,3]. Unaccompanied song tends to have a less regular power distribution.

2. METHOD

Speech and song development and test data were acquired using human subject testing, where subjects produced vocal utterances based on a series of prompts [REFb]. Each datum consists of an acoustic event of approximately 5 seconds. The development data are used to build a computational model of the feature being tested, and the test data are then applied to the model to verify correct classification of *a priori* unseen data.

The models developed in this work relate to the rhythm of an utterance as evidenced by the distribution of silence in the utterance. Utterances are divided into 0.015s frames, and each frame is characterized as silence if the *spl* is below a hysteresis threshold. The threshold is defined by a 5- to 30-frame window at the beginning of the utterances which is assumed to represent the environmental background noise. To allow fuzziness around the threshold, the hysteresis function shown in Figure 1. is used, where T is the *spl* threshold defined by the beginning of the utterance, and t is the allowable deviation. "0" is a silent frame and "1" is a frame with *spl* above the dynamic threshold.

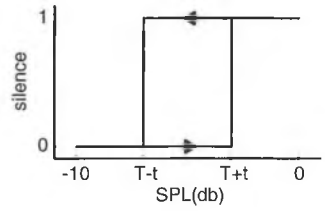

Figure 1. *spl*-to-silence conversion hysteresis function.

3. RESULTS

Silence cues to rhythm can be as simple as the proportion of silence frames in an utterance, which can be used as a feature to discriminate between talking and singing. Figure 2 shows probability distribution estimations for the proportion of silent frames in an utterance, for both talking and singing. Singing files overall tend to have less silence than talking files, although there is considerable overlap.

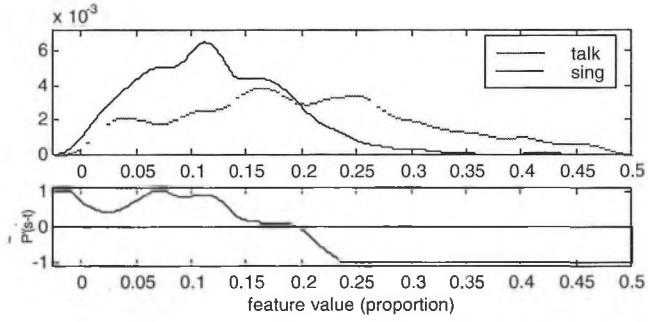

Figure 2. Probability density estimation comparisons for silence.

Rhythmic structure is also examined on a case-by case basis to identify characteristics based on silence patterns. Figures 3 to 7 show examples of these characteristics, which are discussed in Section 4. The figures show *spl* over time, with the upper plot of each figure showing the silence metric generated from the hysteresis function.

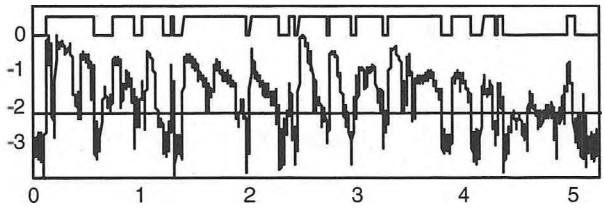

Figure 3. (*l145*) Typical speech.

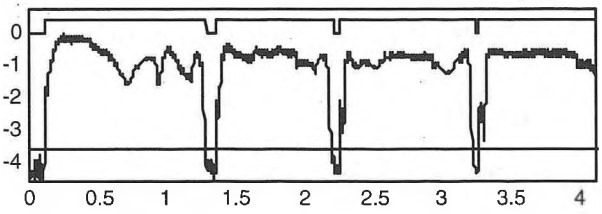

Figure 4. (*m124*) Typical song

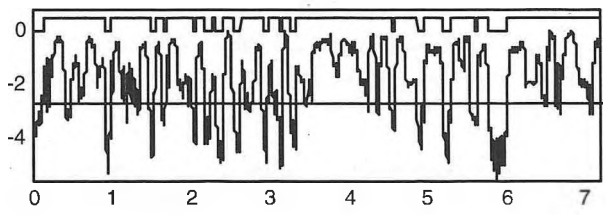

Figure 5. (*m126*) Song with rapid changes

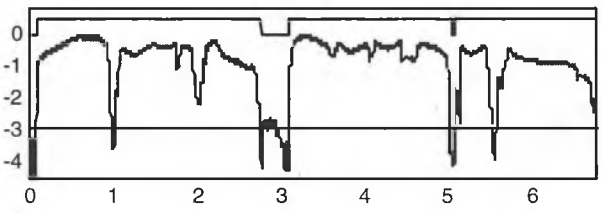

Figure 6. (*g258*) "O Canada...", sung

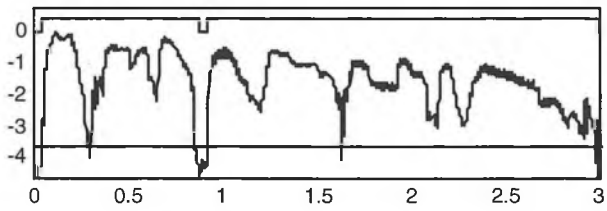

Figure 7. (*h258*) "O Canada...", spoken

4. DISCUSSION

Silence in human speech utterances typically is the result of two situations. When phonetic stops ('p','t','k') are used in speech or song, a small period of silence precedes the plosive. These silences occur during the course of continuous speech, and are not altered in song. The second situation of silence in an utterance is

Figure 3 shows a typical speech utterance. The rhythmic patterns are not related to any specific underlying beat or tempo. Figure 4 shows a typical sung utterance, where a theme is repeated and a series of silence events can be traced across the utterance. The silence events are well spaced and the *spl* track contains a repeating pattern.

Figure 5 shows a sung utterance where the identification of a rhythmic structure is less easy to define. The sung utterance has rapid changes and while there is rhythmic structure, the distribution of silence events is not sufficient to identify an underlying rhythmic structure. As with any analysis of perceptual events, techniques must be moderated with the domain of the perceptual phenomenon.

Figures 6 and 7 are a comparison of speaking and singing utterances of the same lyric by the same subject. This provides a good example of the use of silence as a rhythmic defining structure. The *spl* track shows a period of silence at 2.7s. This period of silence is the space between the first and second phrases of the utterance "O Canada, our home and native land". The spoken version of the utterance has a much smaller gap between the phrases, showing that the longer gap in Figure 6 is primarily for musical emphasis.

Although this paper did not cover the distribution of the *spl* track itself, it is of note that the tracks have significant differences: In the sung track in Figure 6, there is constancy in the *spl* across a syllable, whereas in Figure 7, the corresponding syllables have more variation in the *spl* track.

The rhythmic structure of a musical utterance, therefore, is characterized by regularly spaced silence events representing phonetic stops or phrase boundaries, and the phrase boundary silence events are typically longer than that in speech utterances. This work may be expanded to include the use of *spl* track and silence event analysis to discover rhythmic characteristics in acoustic items other than human utterances., particularly instrumental music.

REFERENCES

[1] List, G. (1971). The Boundaries of Speech and Song. in Readings in Ethnomusicology, D.P. McAllester (ed.). New York, Johnson Reprint Co., 253-268.

[2] Karnebäck, S (2002). Expanded Examinations of a Low Frequency Modulation Feature for Speech/Music Discrimination. Proc of ICSLP2002.

[3] Scheirer, E. and Slaney, M. (1997). Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. IEEE International Conference on Acoustics, Speech and Signal Processing, II:1331-1334.

[4] Gerhard, D. B. (2002). A Human Vocal Utterance Corpus for Perceptual and Acoustic Analysis of Speech, Singing and Intermediate Vocalizations. Journal of the Acoustical Society of America. 112, V: 2264