PERCEPTUAL FEATURES FOR A FUZZY SPEECH-SONG CLASSIFICATION

David B. Gerhard

School of Computing Science, Simon Fraser University
Burnaby, BC CANADA V5A 1S6. email: dbg@cs.sfu.ca

ABSTRACT

Human speech and song seem disparate, but a range of utterances between speech and song are evident, such as poetry, chant, and rap, which have features of both singing and speaking. This work seeks to identify and characterize the perceptual features relevant for a fuzzy classification of utterances between speech and singing. The speech-ness or song-ness of an utterance depends on the speech or song features evident in that utterance. This paper presents a brief discussion of the collection and annotation of the corpus of sound clips used in this work, followed by a description of the perceptual features expected to be useful, and presentation of preliminary results for two of these features.

1. INTRODUCTION

The differences between speech and music have been studied recently, with the aim of developing a system to distinguish between them [1][2]. Such a system could characterize broadcast radio or automatically notate an audio database. The differences between speech and *song* are more subtle because both are monophonic human utterances, and as such share feature values (e.g. bandwidth, energy) which would be different from instrumental or accompanied vocal music. A speech-song classification is useful for improving speech recognition and automatic song transcription, as well as speech therapy and music recognition. Previous work on the discrimination of speech and song [3][4][5] mention features such as pitch, rhythm and vibrato. An assumption of these previous systems is that discrimination consists of choosing between two classes—a clip can only be either speech or song.

Instead, observation of the range of utterances between speech and song indicate that a fuzzy scale between the two classes is more appropriate. Each utterance between speech and song is assigned a "speech-ness" quantity and a "song-ness" quantity, and these combine to produce a fuzzy rating for the utterance. These judgements are perceptual, and as such need a perceptual target. The fuzzy rating for the corpus used in this research was obtained by soliciting listener opinions, as discussed in the following section.

Partially supported by the Natural Sciences and Engineering Research Council of Canada.

2. CORPUS COLLECTION AND ANNOTATION

The corpus of human utterances used throughout this research was collected and annotated using an internet-based form. Phase 1 consisted of subjects providing voice clips in response to prompts, and additional clips were collected from public sources such as radio, television, music and movie soundtracks. A total of 847 clips of about 5 seconds each were collected. Copyright laws were respected [6].

Phase 2 consisted of annotating the clips collected in Phase 1. Clips were separated into three categories: speech, song, and intermediate vocalizations. The intermediate vocalizations were presented on a web page and listeners were invited to rate the clips between speaking and singing, and to provide text feedback about the clips, the speech-song classification as a whole, and the experience of the on-line listener survey. The numeric results are used to provide a target for fuzzy speech-song classification, and the textual results are used to discover perceptual features and characteristics of speech and song that might be useful for classification.

3. FEATURES

The differences between speech and song, and the fuzzy scale between them, are evident in a set of features for which there is a single value for each clip examined. Some features are derived from data reduction methods on the waveform, and some are extracted directly from the clips themselves. The following sections describe potential features and their justification for the fuzzy separation of speech and song.

3.1. Pitch

Pitch is universally identified by non-expert listeners as a principal difference between speech and song. While some listeners find it difficult to describe in what way pitch varies from speech to song, the most obviously useful characteristics of pitch are discreteness and *vibrato*. Sung utterances

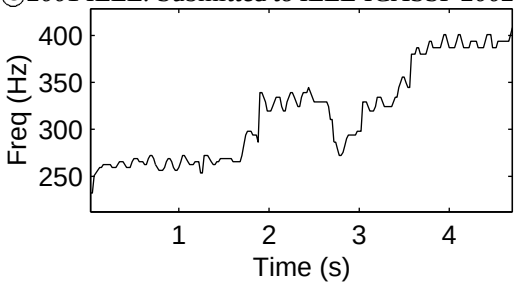

Fig. 1. Example pitch track of a sung utterance.

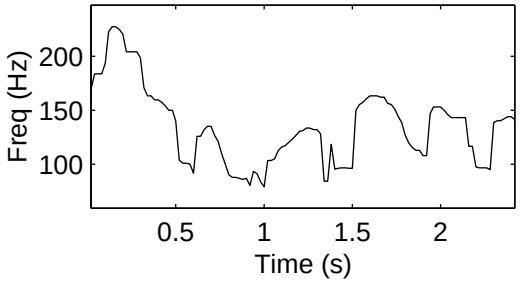

Fig. 2. Example pitch track of a spoken utterance.

often have associated with them a pitch track which oscillates in a characteristic way. Sometimes when people sing they add this oscillation to the pitch of the note they are singing, for stylistic or other reasons. Vibrato is characterized by a stationary pitch augmented by a 6-12 Hz pseudo-sinusoidal waveform.

If a pitch track consists of a series of flat pitch contour sections, separated by discrete jumps, this may be evidence of song, as it could indicate adherence to a musical scale. Similarly, if a pitch track oscillates with a frequency near 6-12 Hz, this could indicate the presence of a vibrato-like effect in the clip, which is evidence of song. Figure 1 shows an example pitch track of a sung utterance with vibrato and discrete pitch levels, and Fig. 2 shows an example pitch track of a spoken utterance.

The pitch track associated with a clip can be found by extracting the fundamental frequency of the waveform at a given frame rate, wherever such a measure is valid. This pitch track can then be distilled into a number of statistics that may be useful, as described in Section 4.

3.2. Voicedness

A preliminary feature for many speech recognition engines is voicedness. Voiced phonemes are produced by a regular glottal pulse from the larynx shaped by the vocal tract, and these utterances have pitch. Unvoiced phonemes (fricatives, plosives, stops, etc.) may have spectral concentrations of energy, but these utterances do not have a pitch *per se*. Voicedness is often used as a pre-processing step for pitch detection: only detect pitch where a pitch is valid. However, when analyzing utterances for a speech-song characterization, voicedness itself is a useful measure.

The proportion of voiced frames in an utterance can be evidence for a speech-song characterization. Often, when people sing, they draw out vowels (voiced) and restrict consonants (usually unvoiced). As a result, sung utterances tend to have a higher proportion of voiced phonemes.

3.3. Rhythm

When asked what makes an utterance speech-like or song-like, many listeners identified rhythm as an important feature, although their descriptions of the specifics of rhythm in the characterization often degrade to circular: speech is characterized by speech-like rhythm and song is characterized by song-like rhythm. To discover what it means for a rhythm to be song-like or speech-like, we must first do a data reduction to extract relevant features for a rhythm measure, and then study the results of this measure on speech clips and song clips.

Preliminary research shows that rhythm in song involves phrase repetition and energy repetition, as well as phoneme repetition. Feature extractors designed to extract repetition, such as autocorrelation and other pitch detection techniques, would be useful for this task.

3.4. Rhyme

Another form of repetition, rhyme is a feature primarily of song, and requires more detailed investigation. Phonetic information would be very useful in detecting patterns of rhyming words, and this would require formant extraction and F1-F2 characterization, similar to the preliminary steps of a speech recognition engine. Rhyme is likely to correlate well with rhythmic structures, so these two features could relate to and inform one another.

3.5. Expectation and Context

Some listeners from the corpus annotation project identified the fact that simple perceptual features may not be sufficient to characterize the fuzzy speech-song axis. Especially in ambiguous utterances, context and expectation play an important role as well. If the lyrics in the utterance are ambiguous, but remind the listener of a song once heard, this recall may be sufficient to nudge the listener's opinion in the direction of song. Similarly, the expected conclusion of an utterance can lend weight to one end or the other of the fuzzy scale. These features are more esoteric and difficult to quantify, but expectational probabilities have been used

in speech recognition engines in the past and could be applied to this problem as well.

4. DESIGN AND PRELIMINARY RESULTS OF FEATURE EXTRACTORS

The design and construction of a fuzzy speech-song classification engine begins with the design and construction of feature extractors relevant to the task. Two relevant feature extractors have been developed at this point, and several sub-features are analyzed from these main features. In this section, preliminary results of the voicedness extractor and the pitch track extractor are presented.

Since the aim of this research is to develop a system capable of making a fuzzy classification between speech and song, classification targets are labeled as speech, song, or intermediate vocalizations, and for a feature to be useful, the intermediate vocalizations should lie between speech and song on the feature axis. Individual features are expected to separate only partially, but as long as they make a minimal separation and present the mean of the intermediate vocalization class between the means of the speech and song classes, the feature is considered to perform well. Feature results will be combined to generate a final rating.

For the purpose of experimentation and design, the corpus is split into two parts: 90% of the clips are put into the "design" corpus and 10% of the clips are put in the "test" corpus. Once the experimentation is complete, the system design will be automated and run 10 times, each time with a different 10% test data, and the final system will be a combination of the resulting 10 classification systems.

4.1. Voicedness

Voicedness is used in two ways. First, to discern where to apply a pitch detector, and second, as a feature in its own right. The energy of a signal is used as a pre-processor for voicedness: if a frame has an energy less than 3 dB above the minimum energy in the signal, it is considered silence. If a frame is silent, voicedness (and hence pitch) is not considered, so the order of examination of these features is:

1. Energy: If a frame is silent, skip other features.
2. Voicedness: If a frame is unvoiced, skip pitch.
3. Pitch.

There are a number of ways to extract voicedness from a clip, two of which are distribution of spectral energy and zero-crossing rate. To extract voicedness using distribution of spectral energy, the following procedure is used: calculate the spectrogram of the signal; for each frame, calculate the energy in the lower frequency bands and consider this the voicedness measure; for each frame, calculate the energy in the higher frequency bands and consider this the unvoicedness measure; compare these two results. If a frame has more energy in the higher frequency bands, it is labeled as unvoiced, and if it has more energy in the lower frequency bands, it is labeled as voiced. This procedure has complexity $O(n \log n)$ for the FFT, and $O(n)$ to calculate the frame energies, for a total complexity of $O(n \log n + n)$.

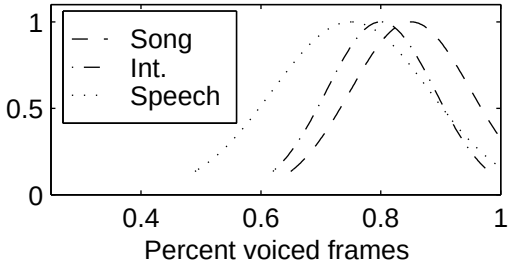

Fig. 3. Category separation: proportion of voiced frames.

The zero-crossing rate can be used as a voicedness measure because fricatives (the majority of unvoiced frames) have a higher zero-crossing rate than do voiced frames [7]. The procedure is this: for each frame in the clip, calculate the number of times the signal crosses zero. This can be done in hardware, or a software implementation has complexity $O(n)$. If a frame has a zero-crossing rate above a statistically determined threshold, it is considered unvoiced, otherwise it is considered voiced. The third possibility for a frame is silence, and so a voiced frame is a frame which is neither unvoiced nor silent.

The ratio of voiced frames to total frames is measured for each clip in the design corpus and the results are presented in Fig. 3. The utterances are pre-labeled as speech, song, and intermediate vocalizations, and the best fit normal distributions of each category are plotted. The voicedness feature presents the mean of the intermediate vocalizations between the means of the speech utterances and song utterances, and so is considered successful.

4.2. Pitch

The pitch track $f_0(t)$ for each clip in the design corpus is calculated using the autocorrelation method, and is considered valid only for voiced frames. A pitch measure is calculated for each frame, and features are extracted from the resulting pitch track. The cepstrum method was also considered, although results for the autocorrelation pitch extraction algorithm were consistently better than for the cepstrum method.

For the speech-song characterization, there are several potentially relevant statistical features of pitch. Song tends to have a wider pitch range, although a narrow pitch range

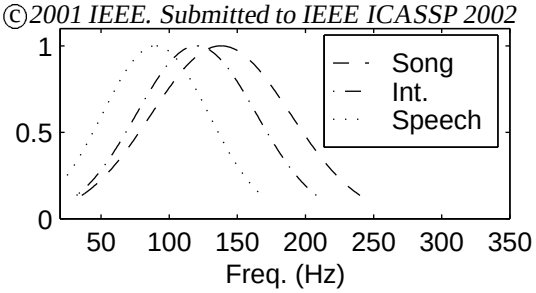

Fig. 4. Category separation: minimum pitch.

does not necessarily indicate speech. Song tends to have larger extremes, although more restricted extremes do not necessarily indicate speech. As with all of these features, individually they may not provide much information, but in combination they can be very powerful. Since song often contains pitches higher than in normal speech, and often uses a wider pitch range, we might expect minimum pitch, maximum pitch, mean pitch and/or pitch range to be useful in distinguishing speech from song. For continuity, pitch range is calculated as

$$R_{f_0(t)} = \frac{MAX(f_0(t)) - MIN(f_0(t))}{\sigma_{f_0(t)}}, \qquad (1)$$

where $\sigma_{f_0(t)}$ is the standard deviation of the pitch track $f_0(t)$. It is surprising to find that of these measures, preliminary results show that minimum pitch is the best overall at separating the categories in the design corpus. The normal distribution of minimum pitch across the design corpus for the speech, song, and intermediate vocalizations is presented in Fig. 4. The mean minimum pitch for intermediate vocalizations is between the mean minimum pitches for speech and song, and so minimum pitch is considered a successful feature.

As discussed above, the pitch track is also potentially useful for calculating more complicated features. Discreteness and vibrato are features which are relatively straightforward to calculate, and appear to be potentially useful in the speech-song characterization. The pseudo-sinusoidal waveform present in vibrato-like utterances can be detected using spectral analysis of the pitch track. Spectral analysis of the *derivative* of the pitch track may be more effective, because pitch shifts and other irregular modulations will be removed, but the sinusoidal component will be retained because the derivative of a sinusoid is itself a sinusoid.

Pitch constancy is another characteristic that is likely to be evidence of song as opposed to speech. When people sing, they tend to stay on a note for a few tenths of a second before moving to another note, and staying there. In speech, the frequency changes in a more continuous way, sweeping across frequencies. It is expected that some form of statistical stationarity detector will be able to measure the discreteness, and provide evidence for song. In this case, it would clearly be desirable to first detect and remove any vibrato from the signal. This is another example where features interacting with each other could provide a better judgement than combining the ratings from separate features.

5. CONCLUSIONS

Listener studies and personal observation show that there is a range of human utterances between speech and song, including counting rhymes, dramatic reading and storytelling, and that a fuzzy classification of these utterances is appropriate. This classification is perceptual, so perceptual features are investigated. Listener ratings from a corpus collected and annotated by the author provide classification targets. Many perceptual features, such as pitch, rhythm, and expectation, are potentially useful for performing such a fuzzy classification. Preliminary results for voicedness, as well as several features derived from the pitch track indicate that these features, when combined with others, will contribute to a workable fuzzy classification system.

6. REFERENCES

[1] John Saunders, "Real-time discrimination of broadcast speech/music," in *"International Conference on Acoustics, Speech and Signal Processing"*. IEEE, 1996, pp. 993–996.

[2] Eric Scheirer and Malcolm Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *"International Conference on Acoustics, Speech and Signal Processing"*. IEEE, 1997, vol. II, pp. 1331–1334.

[3] George List, "The boundaries of speech and song," in *Readings in Ethnomusicology*, D.P. McAllester, Ed., pp. 253–268. Johnson Reprint Co., 1971.

[4] Esther Ho Shun Mang, *Speech, Song and Intermediate Vocalizations: A Longitudinal Study of Preschool Children's Vocal Development*, Ph.D. thesis, University of British Columbia, 1999.

[5] Tong Zhang and C.-C. Jay Kuo, "Heuristic approach for generic audio data segmentation and annotation," in *Proceedings of ACM International Multimedia Conference*, November 1999.

[6] United States Code, "Title 17: Copyright, Section 107: Limitations on exclusive rights: fair use," 2000.

[7] Benjamin Kedem, "Spectral analysis and discrimination by zero-crossings," *Proceedings of the IEEE*, vol. 74, no. 11, pp. 1477–1493, Nov. 1986.