Pitch-Based Acoustic Feature Analysis for the Discrimination of Speech and Monophonic Singing

David Gerhard
Computing Science, Simon Fraser University, 8888 University Dr., Burnaby, BC Canada, V5A 1S6. dbg@cs.sfu.ca

1. INTRODUCTION

A system capable of discriminating between human speech and human monophonic (*solo a capella*) singing would be a useful tool for several classes of applications including query-by-humming and content-based annotation and search of multimedia databases, as well as speech and music therapy and music education. Differences between speech and song have been investigated by List [1], and previous work by Zhang [2] contains a section on discriminating between speech and song (both with background music) in the context of a general audio classification scheme, using three features to make this classification: harmonic ripple, harmonic segment duration, and fundamental frequency above 300 Hz. Eric Scheirer and Malcom Slaney [3] have done work on discriminating speech and music using a set of 13 features, some of which are applicable to the speaking versus singing task described in this paper.

To investigate the perceptual differences between talking and singing, human subjects were exposed to a corpus of singing and talking sounds, and asked first to classify each sound on a scale between speaking and singing, and then to indicate the characteristics of the sounds that lead to their judgements. The subject responses indicated that pitch is a primary factor in making this judgement. Subjects indicated many pitch-based subfeatures including vibrato (similar to Zhang's harmonic ripple), excessively low or high pitch, adherence to a musical scale, and smoothness of pitch. Other features not directly related to pitch include rhythm, rhyme, context and expectation. As will be seen later in this paper, some of these features can also be investigated using pitch as a base feature.

2. METHOD

To develop a classification engine for the talking versus singing discrimination task, feature extractors are constructed based on perceptual characteristics indicated in the human subject trials. The full classification engine takes as input a sound file, extracts base features from this file, extracts subfeatures from these base features, combines these features using dimensionality reduction, and then presents a classification based on the nearest representative class in the experience of the classifier. This paper presents the extraction models for features based on the fundamental frequency (f_0), the physical counterpart of pitch.

2.1 f_0 Extraction

For this work, several current f_0 extractors were considered based on periodicity detection algorithms such as cepstrum and autocorrelation. The method used in this work is an autocorrelation-based algorithm [4] with several domain-specific improvements. f_0 measurements were extracted from each sound file at 15ms intervals. Example f_0 tracks are presented in Figures 1 and 2. These f_0 tracks are used as a base for the subfeatures described in Section 2.2.

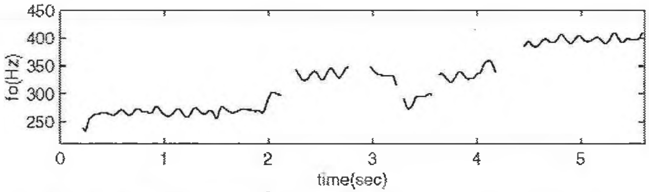

Fig. 1. Example sung f_0 track: "Row, row, row your boat."

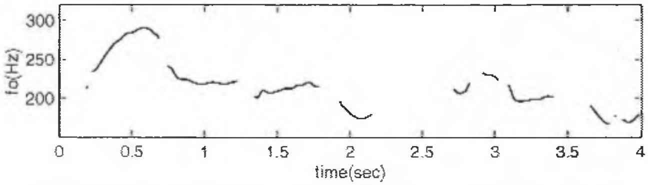

Fig. 2. Example spoken f_0 track: "Row, row, row your boat."

2.2 Subfeatures of f_0

The characteristics indicated in the human subject trials described in Section 1 inspire a set of f_0 subfeatures. Most of these subfeatures are direct derivations from the f_0 track, while some use additional information.

Vibrato

The most obvious difference between the f_0 tracks in Figures 1 and 2 is the presence of a ripple in the sung utterance. This phenomenon is caused by the singer modulating his or her vocal pitch [5] using a modulation frequency near 4 - 8 Hz. Vibrato is not present in all sung utterances, but it is present in very few spoken utterances. Vibrato indicators are extracted from the f_0 track using two standard periodicity detectors: autocorrelation and cepstrum.

f_0 Statistics

Many subjects commented on how sung utterances were higher or lower or more consistent in pitch. First order statistics are extracted from the f_0 track to measure these perceptual differences. Four statistical features are extracted: max, min, mean (μ) and standard deviation (σ). The slope of the f_0 track is also expected to be informative, indicating the rate of change of the pitch. The same statistical measures are used for the f_0 track slope.

Correlation between Syllables

Phrase repetition is a common characteristic of sung utterances. Often, a segment of pitch will be repeated with the same or different words, and this phenomenon is uncommon in spoken utterances. The correlation (R) between syllables is one way to identify the presence of this phenomenon. The f_0 track is separated into syllables at non-pitched segments and segments of high pitch slope. The set of f_0 segments ($f_0(k)$) is then correlated to obtain a measure of the similarity between syllables in the utterance.

3. RESULTS

Each feature described in Section 2 is implemented and tested on a corpus of singing and talking samples. The feature models are generated using the following procedure: Each feature extractor is presented to the set of singing files and the set of speaking files, and a statistical profile is determined based on the estimated likelihood of each feature value. These estimated probability density functions (*pdf*s) are then combined to generate the final feature model: where the *pdf* of the speech files is greater than the *pdf* of the song files, the feature values are evidence for speech, and vice versa. If the difference between the *pdf*s is below a threshold, the feature value is neutral.

Two feature models are presented here as examples. The "class" plot indicates the classification at each feature value. 1 = speaking, 5 = singing, 3 = no evidence.

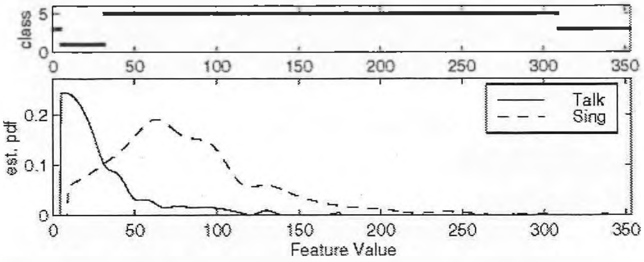

Fig. 3. Feature model of autocorrelation-based vibrato.

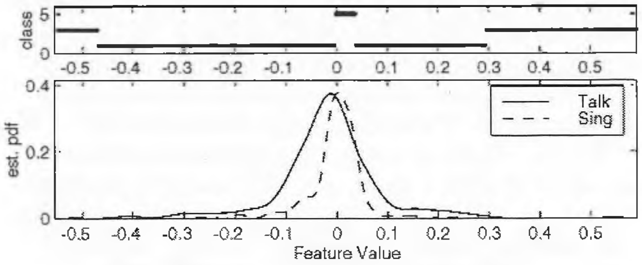

Fig. 4. Feature model of mean pitch slope.

Each feature model is evaluated based on the number of corpus files correctly classified. The correct rate for each feature is presented in Table 1.

Table 1. Percent correct for f_0-based features.

Feature	Rate	Feature	Rate
Vib AC	69.78%	max(f_0')	55.93%
Vib Cep	72.12%	min(f_0')	63.94%
max(f_0)	65.28%	μ (f_0')	65.28%
min(f_0)	54.09%	σ(f_0')	47.41%
μ (f_0)	68.78%	R ($f_0(k)$)	78.63%
σ(f_0)	55.93%		

4. DISCUSSION

Because some files do not exhibit the phenomena tested for in each feature, some individual features perform poorly. The next stage in this work is to develop multi-feature models. Dimensionality reduction is expected to be successful - because the features presented here are all based on a single superfeature, it is likely that some of these features are measuring the same underlying phenomenon.

A pitch-based feature that would seem useful in the speech/song task is pitch continuity. A song (it seems) is a series of discrete pitches, and it should be a simple task to recognize this in the pitch track. The difficulty with this is twofold: the human oratory system is not good at generating a stationary pitch, and the human auditory system is very good at recognizing non-stationary pitch tracks as consistent notes, using mechanisms which are not currently understood. Even identifying the intended target pitch of an utterance *known* to be song remains a difficult task.

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] Zhang, T and Kuo, C.-C. J. (1999). Heuristic Approach for Generic Audio Data Segmentation and Annotation. Proceedings of ACM International Multimedia Conference.

[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] de Cheveigné, A and Kawahara, H. (2002). YIN, a Fundamental Frequency Estimator for Speech and Music. Journal of the Acoustical Society of America, 111: #4.

[5] Seashore, C. E. (1936). Psychology of the Vibrato in Voice and Instrument. Iowa: The University Press.

AUTHOR NOTES

The work was conducted while Mr. Gerhard was a student at Simon Fraser University.