MUSIC TONALITY FEATURES FOR SPEECH/MUSIC DISCRIMINATION

Gregory Sell and Pascal Clark

Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD, USA
{gsell,pclark}@jhu.edu

ABSTRACT

We introduce a novel set of features for speech/music discrimination derived from chroma vectors, a feature that represents musical tonality. These features are shown to outperform other commonly used features in multiple conditions and corpora. Even when trained on mismatched data, the new features perform well on their own and also combine with existing features for further improvement. We report 97.1% precision on speech and 93.0% precision on music for the Broadcast News corpus using a simple classifier trained on a mismatched corpus.

Index Terms— voice activity detection, music detection, amplitude modulation, chroma

1. INTRODUCTION

Speech/music discrimination or classification is highly relevant to both speech and music processing. In some cases, such as monitoring radio broadcasts for content type, the classification and annotation is the sole goal. In other cases, however, distinguishing between music and speech is a front-end to a downstream application, whether it be automatic speech recognition or music genre classification.

Segmenting an audio stream prior to a costly process is valuable for several reasons. For one, recognizing sections of a signal that do not pertain to the task at hand reduces computation time and allows for more efficient resource allocation. Additionally, front-end audio classification can purify data for more accurate training models and cleaner testing. These issues are especially important when the data contains both speech and music, because speech and music are both rich signals that are often difficult to distinguish for generic activity detectors.

Speech/music classification research has examined a variety of techniques in the last several decades. Saunders [1] used several statistics of zero-crossing rate (ZCR) to label FM radio broadcasts. Scheirer and Slaney [2] explored a larger set of features with several classifiers and found that syllabic-rate amplitude modulation energy, spectral flux variance, and overlaps in multi-band autocorrelations (called the pulse metric) were the most effective combination. Another study [3] focused on single-feature detectors and found cepstral features to be best, followed by amplitude, pitch, and then zero-crossing rate. In [4], linear discriminant analysis was applied to a large feature set with success. Pangiotakis and Tziritas [5] used a sieve-like approach, which is created by connecting a series of high precision detectors that apply pre-determined thresholds to low dimensional features like ZCR and root-mean-square (RMS) variance.

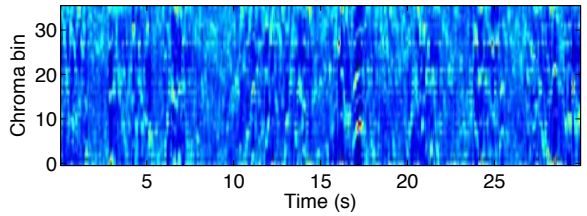
The goals and constraints of speech/music discrimination are highly task specific. Some applications can afford high-complexity solutions, such as the the multi-layer perceptron and HMM classifiers in [6] or [7]. In [8], both a sieve-like structure and a learning algorithm are placed in series. Alternatively, in several of the previously mentioned studies, such as [1, 2, 5], fast and efficient computation is a top priority, and so some accuracy is sacrificed in the name of speed.

In this paper, we are interested in the set of applications that require a lightweight computation. We will introduce a new feature set for speech/music classification based on *pitch class profiles* or *chroma vectors*, a feature typically used in music tonality tasks such as chord or key identification. We will then compare this new feature to several of the best features from previous research using simple classifiers on the publicly available GTZAN Music/Speech dataset [9]. Finally, we will demonstrate the cross-corpus applicability of these features by annotating speech and music in the Broadcast News corpus [10], treating it strictly as a held-out evaluation data set.

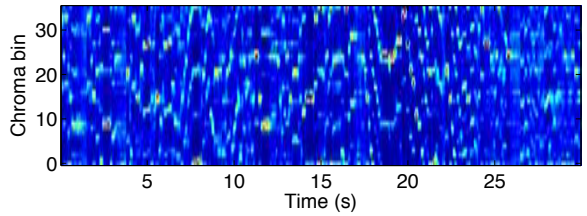
2. FEATURES FOR SPEECH/MUSIC DISCRIMINATION

Past research has examined many features for speech/music discrimination, but the most successful features are designed to exploit several known differences between the structure of speech and music. Examples of characteristics utilized in prior work include the alternation in speech between voiced and unvoiced sections, or the wideband attack common in musical percussion. In this section, we introduce a new discrimination feature that exploits the lack of musical scale in speech. Music, on the other hand, is typically dictated by specific keys and tonal structures that follow strict patterns in the frequency

978-1-4799-2893-4/14/$31.00 ©2014 IEEE



(a) Speech Chromagram



(b) Music Chromagram

Fig. 1. Example chromagrams for speech and music samples. Note the greater prevalence of peaks in music chroma as compared to speech.

domain. Saunders mentioned a similar concept in [1] but, to our knowledge, the use of music tonality features has not been attempted until now.

Pitch-class profiles [11], or chroma vectors, are used in music processing for any task that involves the tonality of music, such as key or chord identification. They utilize a principle known as *octave invariance* which states that there is no functional difference between musical notes separated by a doubling of frequency. In other words, a chord in one octave serves the same musical purpose as the same chord in another octave. Pitch-class profiles utilize this principle to reduce the spectrum $X(f)$ by summing exponentially separated frequencies into the same bin, essentially folding octaves of the spectrogram into the same range.

$$\text{chroma}(k) = \sum_{r=0}^{R-1} |X(2^{\frac{k+rK}{K}} f_{min})| \qquad (1)$$

Here, we are calculating the k^{th} chroma bin, and K is the total number of bins in the chroma vector, R is the number of octaves spanned by the computation, and f_{min} is the lowest frequency included in the summation.

However, this chroma feature alone is not particularly effective for speech/music classification. The chroma feature captures information about chord and key which will vary greatly between musical segments, leading to a highly modal distribution of all music. For the purpose of speech/music discrimination, we propose a simple, yet powerful, measure of peakiness in the chroma vector.

Musical tones will tend to locate around certain frequencies more often than others. The commonality of these frequencies and their relationships are determined by the music theory of that particular culture or style of music, but almost all music follows some basic set of rules regarding notes and relationships between those notes. Speech, on the other hand, is far less strictly regulated in the use of pitch. These differences will lead to stronger and more separated peaks in the chroma vectors of music, while the chroma vectors of speech will tend to be smoother with mounds of energy around the bins corresponding to formant and fundamental frequencies of the speech. As a result, we expect musical chroma vectors to be more peaked as a function of k when compared to speech chroma vectors. This characteristic is clear in the examples in Fig. 1.

We explore two metrics of this characteristic. For one, we calculate the energy after differentiation of a normalized chroma vector.

$$\text{Chroma Diff.} = \sum_{k=0}^{K-1} |c(k) - c(\text{mod}(k+1, K))|^2 \qquad (2)$$

Here, c is the chroma vector from Eq. (1) after energy normalization. Note that the differentiation is calculated circularly (with modulo function), because the musical tones represented by the chroma vector are circularly related.

The second proposed feature is calculated by summing the high-frequency energy in the normalized spectrum.

$$\text{Chroma High Freq.} = \sum_{l=l_{min}}^{l_{max}} |\mathcal{F}\{c(k)\}(l)|^2 \qquad (3)$$

Here, $\mathcal{F}\{\cdot\}$ represents the Fourier transform, and so the feature is the total energy in the spectral range [l_{min},l_{max}].

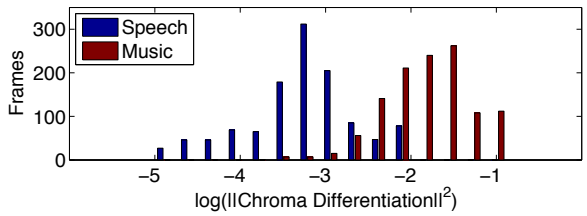
The histograms for these two proposed features (corresponding to the chromagrams in Fig. 1) are plotted in Fig. 2, showing clearly separate speech and music distributions with only small overlap. Note that log-compressed features are used to gaussianize the energy metrics.

The specific parameters used to calculate these features will be discussed in Section 3.2.

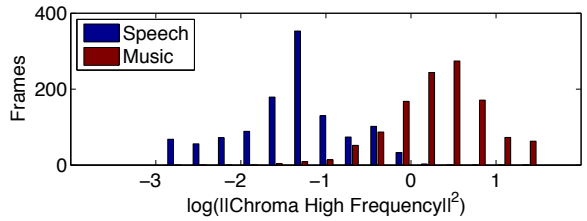
3. EXPERIMENT

We tested the value of these new choma-derived features in two experiments. First, we ran a preliminary evaluation and feature selection on the GTZAN Music/Speech dataset, which is composed of 64 examples of 30-second speech recordings and 64 examples of 30-second music recordings. This dataset is diverse, containing multiple styles of music as well as speech recorded in many conditions. For previous studies using this corpus, refer to [12, 13].

Our secondary goal was to examine the cross-corpus compatability of these features, so we used models trained on the GTZAN Music/Speech data to discriminate speech and music



(a) Chroma Differentiation Histogram



(b) Chroma HFE histogram

Fig. 2. Histograms for the novel features derived from chroma representations. In each case, the greater peakiness in music chroma results in larger feature values.

segments from the Broadcast News corpus [10]. In this experiment, we only examined sections that were labeled as speech or music (ignoring segments that contain neither). Broadcast News differs from the GTZAN corpus in that speech and music both appear in the same file and with arbitrary duration, and also because speech and music are imbalanced in the set (90.5% speech, 2.4% music, 7.1% speech and music).

Both corpora were downsampled to 8 kHz prior to feature computation.

One shortcoming of these experiments is that they only test non-tonal English. Examining the efficacy of these features on tonal languages (which may behave differently in the chroma space) would be a valuable task for the future.

3.1. Features

In addition to the chroma-derived features, we computed a collection of features that have been successful in past research. Short-time metrics (such as zero-crossing rate and spectral centroid) were calculated for 25ms windows every 10ms, after which their statistics were calculated over 1 second frames.

- **Normalized RMS standard deviation** - Standard deviation of short-time RMS divided by the mean. [5]
- **Silent interval ratio** - The proportion of short-time RMS values that are below the mean for the frame. [5]
- **Silent interval frequency** - The number of continuous segments of short-time RMS measurements that are below the mean RMS for the frame. [5]
- **ZCR variance** - Variance in the zero-crossing rate within short-time frames. [1]
- **Spectral centroid variance** - Variance in the spectral centroid of short-time spectra. [2]
- **Spectral flux variance** - Variance of the energy in the differentiation of neighboring short-time spectra. [2]
- **Mel-frequency subband modulation syllabic rate energy** - Total energy near 4 Hz in the modulation spectra of mel-frequency subbands. [2]
- **Mel-frequency subband modulation spectral centroid** - Spectral centroid in the average of modulation spectra calculated on mel-frequency subbands (a feature used in speech activity detection). [14]

Tempo-based features have also been used in the past (such as the pulse metric in [2]), but we found such measures are not as effective when limited to 1 second of audio context.

3.2. Chroma Feature Parameter Selection

There are a few parameters for the new features that require selection. First, we must select the number of chroma bins (K in Eq. (1)). Typically, 12 bins are used, because there are 12 Western music pitch classes. However, in order to encourage a greater degree of peakiness, and also to generalize beyond Western music, we examined using multiples of 12. After preliminary experimentation, we found performance improved steadily with increasing number of bins, before saturating around 36. We also set f_{min} in Eq. (1) to 220Hz (a common tuning for A3), and, for the high-frequency feature (Eq. (3)), we set the summation boundary bins l_{min} and l_{max} to 6 and 18, respectively.

Both types of chroma features were calculated for 100ms windows with a 25ms hop, then averaged over one second frames prior to the calculations in Eqs. (2) or (3).

3.3. Results

The results are separated into three separate sets. First, we examined all 10 features as individual detectors. Then, we performed feature selection to find the best multi-dimensional classifier. Both of these experiments were performed on the GTZAN Music/Speech dataset and fit a Gaussian distribution to each type of signal. Frames were assigned to the model with the higher likelihood. Evaluations were performed using 8-fold cross-validation with separate files in train and test datasets.

Third, we conducted a domain-mismatch experiment using the system resulting from the feature selection process. The new domain for our experiment was the Broadcast News database, which consists of audio from television and radio shows. We annotated the HUB4 set using our system trained

GTZAN Music/Speech	Overall	Speech	Music
Norm. RMS StDev	82.1%	85.4%	78.9%
Silent Interval Freq.	75.3%	66.1%	84.5%
Silent Interval Ratio	72.6%	95.2%	50.0%
Norm. ZCR Var.	76.6%	80.8%	72.5%
Spec. Centroid Var.	81.1%	84.2%	78.0%
Norm. Spec. Flux Var.	74.2%	75.6%	72.8%
Mod. Spec. Syllabic	77.6%	75.8%	79.5%
Mod. Spec. Centroid	79.2%	81.3%	77.1%
Chroma First Diff.	86.2%	87.1%	85.2%
Chroma High Freq.	**86.6%**	87.6%	85.6%

Table 1. Overall classification accuracy and recall results for 8-fold cross validation on the GTZAN Music/Speech dataset.

Feature Selection	Overall	Speech	Music
Chroma High Freq.	86.6%	87.6%	85.6%
+ Mod. Spec. Centroid	91.9%	93.5%	90.2%
+ Norm. RMS StDev	93.5 %	95.2%	91.7%

Table 2. Results for the first three features selected using the GTZAN Music/Speech dataset.

on the GTZAN corpus in order to determine how our features adapt to data from an unseen domain.

As part of the domain-mismatch experiment, we also tested a baseline system using generic features. Our baseline used mel-frequency cepstral coefficients (MFCCs) with deltas and double-deltas (calculated with RastaMat [15]) in separate speech and music GMMs with 16 components each (also trained exclusively on the GTZAN corpus before testing on the held-out Broadcast News audio).

3.3.1. Single-Feature Detectors

Results for the single-feature detectors are shown in Table 1. It is immediately evident that both new features yield the best accuracies. The chroma high-frequency feature slightly outperforms chroma differentiation, but it is worth also noting that chroma differentiation could potentially be preferred for its slightly faster computation (requiring only a first-order difference instead of a second FFT). Other features, such as normalized RMS standard deviation, spectral centroid variance, and modulation spectral centroid, are also effective, but the best accuracies are achieved with the chroma features.

3.3.2. Feature Selection

The feature selection process for the first three features is shown in Table 2. The chroma high-frequency energy is the first selected feature. The modulation spectral centroid provides the biggest additive gain to chroma high-frequency energy, even though several other features are individually better. The next feature selected was normalized RMS standard deviation, after which performance saturates. These three features combine to make a quick and lightweight detector that is able to successfully classify 93.5% of the speech/music audio.

Broadcast News Test	Speech	Music
MFCCs w/ 16-GMM	88.2%	89.6%
Chroma High Freq. w/ 1-GMM	84.2%	90.3%
Mod. Spec. Centroid w/ 1-GMM	91.1%	76.0%
Norm. RMS StDev w/ 1-GMM	93.7%	76.4%
Best 3 w/ 1-GMM	**97.1%**	**93.0%**

Table 3. Test results on the Broadcast News dataset using models trained on the GTZAN Music/Speech dataset.

3.3.3. Channel Mismatch

The results on Broadcast News for the selected features and baseline MFCC system are shown in Table 3. These precision rates are calculated for speech without music or music without speech, as the desired classification for overlapping regions depends on the downstream application (and can be controlled by adjusting the decision threshold).

Each single-feature classification performs reasonably well (especially considering the complexity of the task and simplicity of the model), but the combination of the three yields very high precision, despite the mismatch between train and test data. Also, the combined system outperforms the MFCC baseline (also trained on GTZAN Music/Speech) for both speech and music.

These numbers could likely be further improved by incorporating temporal context, either by filtering of the detection scores or a higher complexity method like an HMM.

4. CONCLUSION

We introduced a new set of chroma-based features for classifying speech and music and demonstrated that they improve on the best features for the task found in past research. The chroma features are quick to calculate and effective, yielding high accuracy rates even on their own with a simple classifier. However, when paired with two other single-dimensional features, speech in the Broadcast News dataset was classified with 97.1% precision, and music with 93.0% precision. This is especially noteworthy as it was performed on real-world audio after training on completely mismatched data. The experiments demonstrate that these simple chroma or tonality features create a lightweight but robust system for distinguishing between speech and music. In the future, we would like to examine the effect of tonal languages on the separability of speech and music using these features.

5. REFERENCES

[1] John Saunders, "Real-Time Discrimination of Broadcast Speech/Music," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 1996.

[2] Eric Scheirer and Malcolm Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 1997.

[3] Michael J. Carey, Eluned S. Parris, and Harvey Lloyd-Thomas, "A Comparison of Features for Speech, Music Discrimination," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[4] Enrique Alexandre, Manuael Rosa, Lucas Cuadra, and Roberto Gil-Pita, "Application of Fisher Linear Discriminant Analysis to Speech/Music Classification," in *Proceedings of the 120th Audio Engineering Society Convention*, 2006.

[5] Costas Panagiotakis and George Tziritas, "A Speech/Music Discriminator Based on RMS and Zero-Crossings," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 1–12, February 2005.

[6] Jitendra Ajmera, Iain McCowan, and Hervé Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," *Speech Communication*, vol. 40, pp. 351–63, 2003.

[7] Gethin Williams and Daniel P. W. Ellis, "Speech/Music Discrimination Based on Posterior Probability Features," in *Proceedings of Eurospeech*, 1999.

[8] Yizhar Lavner and Dima Ruinskiy, "A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation," *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[9] George Tzanetakis, "Gtzan_musicspeech," availabe online at http://marsyas.info/download/data_sets/, 1999.

[10] David Graff, John Garofolo, Jonathan Fiscus, William Fisher, and David Pallett, "1996 English Broadcast News Speech (HUB4)," Linguistic Data Consortium, Philadelphia, 1997.

[11] Takuya Fujishima, "Realtime Chord Recognition of Musical Sound: a System Using Common Lisp Music," in *Proceedings of ICMC*, 1999.

[12] George Tzanetakis and Perry Cook, "A framework for audio analysis based on classification and temporal segmentation," in *Proceedings of EUROMICRO*, 1999.

[13] George Tzanetakis and Perry Cook, "Sound Analysis Using MPEG Compressed Audio," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 2000.

[14] David C. Smith, Jeffrey Townsend, Douglas J. Nelson, and Dan Richman, "A Multivariate Speech Activity Detector Based on the Syllable Rate," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, 1999.

[15] Daniel P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," available online at http://labrosa.ee.columbia.edu/matlab/rastamat/, 2005.