Speech to Song Classification

Emily Graber*
Center for Computer Research in Music and Acoustics, Stanford University
*contact emgraber@stanford.edu

Aims and Motivations

~Understand the neural mechanism governing the perceptual classification of Speech to Song (STS) stimuli
~Investigate possible features to predict the STS illusion.
~Compare classification performance using spectral features, linguistic features and music-theoretical features

Speech to Song Stimuli

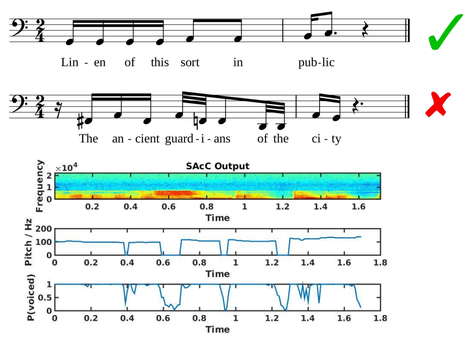


Figure 1. Top and bottom staves show the transcription of a transforming and non-transforming clip respectively. Lowest plots show the spectrogram, the pitch and voiced onsets estimated by SAcC for the transforming stimulus (top staff). The text is 'Linen of this sort in public'.

The STS illusion occurs when a short excerpt of regular speech is repeated about 10 times. The speech in the clip is perceived to transform into a clear melody, or song. Not all clips can transform however. Brain imaging studies have shown reliable neural correlates that differentiate the two types of stimuli, yet the acoustic/semantic reasons for the differentiation are not known.

48 clips with mean duration 1.3859 seconds (SD = 0.3923) were excerpted from audiobooks[1]. Clips are matched for syllable length, syllable rate, and average f_0 (See Tierney et al. 2013). The clips contain only speech, but some of the speech can be perceived as song after repeated hearings.

Raw data was processed from mono wavefiles (fs = 44100) into 13 dimensional mel-frequency cepstral coefficients (MFCCs) with ~20 ms Hann windows, and 50% overlap[2]. Segmentation, i.e. voiced onsets and pitches were estimated by Subband Autocorrelation Classification (SAcC)[3]. Additional features were derived after transcribing the audio.

(Logistic Regression) and CART

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$p(y \mid x; \theta) = (h_\theta(x))^y \, (1 - h_\theta(x))^{1-y}$$

$$\ell(\theta) \;=\; \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))$$

The hypothesis above is a sigmoid curve. If the labels, *y*, take on the values 1 or 0, then the probability of a certain class given the data is expressed as *h(x)* or *(1 - h(x))*. The total probability given every training sample is a product of the probabilities, thus the log-likelihood reduces into a sum over the training samples of *log(p(y | x; θ))*. Estimates of the parameters are found by maximizing the log-likelihood.

Classification and Regression Trees (CART) have been utilized for their interpretability. Logistic regression is useful for capturing the fluidity between perceptual categorical boundaries.

Feature Categories

General: key, number of notes, number of unique notes, total duration

Language: number of syllables, number of stressed syllables, number of syllables in the longest word

Spectral: averaged MFCCs

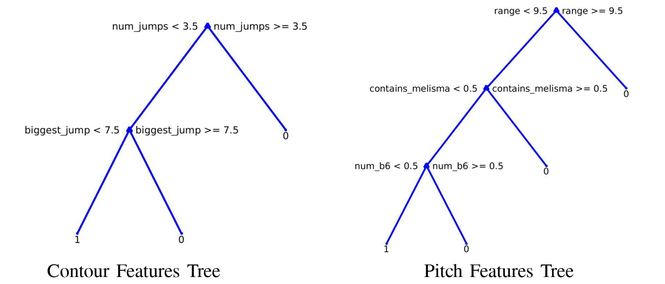
Pitch: histogram of scale degrees, range in semitones, melisma

Contour: number of melodic leaps and steps, largest leap size in semitones, number of consecutive leaps

Harmonic: implied tonic, dominant or subdominant, mode, non-diatonic pitches, resolution level and strength

Rhythmic: total number of onsets, number of strong beats, pickups, syncopations, hemiolas, implied meter

Results



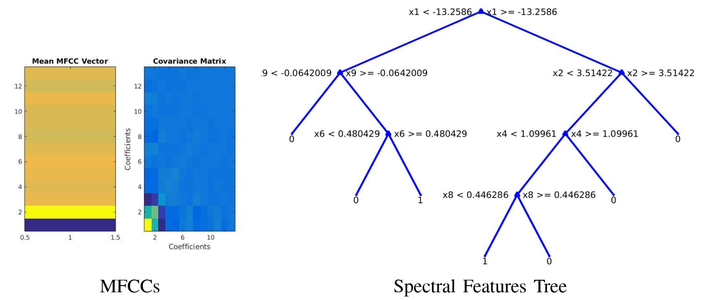
Contour Features Tree Pitch Features Tree

Figure 2. CART results. Left tree shows classification based on contour features, and right tree shows classification based on pitch features.

	LOOCV test error	hit rate	miss rate	false alarm rate	correct rejection rate
General	0.4167	0.5833	0.4167	0.4167	0.5833
Language	0.7292	0.25	0.75	0.7083	0.2917
Pitch	0.2708	0.6667	0.3333	0.2083	0.7917
Contour	0.125	0.875	0.125	0.125	0.875
Harmonic	0.3125	0.7917	0.2083	0.4167	0.5833
Rhythm	0.5	0.5833	0.4167	0.5833	0.4167
All	0.16667	0.875	0.125	0.2083	0.7917

Table 1. CART results. Contour features outperform classification when all features are available. Hit counted when the stimulus turns to song and the prediction is song. Miss counted when the stimulus turns to song and prediction is not song. False alarm counted when the stimulus is not song yet song is predicted. Correct rejection is counted when the stimulus is not song and the prediction is not song.

Results Continued



MFCCs Spectral Features Tree

Figure 3. CART results with spectral features (shown on the left). Each split is based on the value of certain cepstral coefficients.

Discussion

Though many audio machine learning algorithms make use of spectral features, or distributions of spectral features to classify audio, this application introduces a unique dataset for classification where both classes of audio would, under normal circumstances, be called clean speech. As shown above, spectral features cannot be used to easily differentiate between those STS clips that transform into song and those that do not.

Based on the data shown in table 1, the feature set which best classifies the data is the melodic contour feature set. The pitch feature set follows closely, as does the harmonic feature set.

This suggests that our perceptual categorization of the STS clips is closely tied to inherent tonal aspects of the clips. In general 'good' melodies tend to have smooth contours (see root of the contour tree). Melodies that are easy to produce also tend to have smaller ranges (see root of pitch tree). The role of speaking now needs to be distinguished from the role of tonal music theory in this perceptual phenomenon.

References

[1] Tierney, A., Dick, F., Deutsch, D., and Sereno, M. (2013). Speech versus Song : Multiple Pitch-Sensitive Areas Revealed by a Naturally Occurring Musical Illusion. Cerebral Cortex, 23:249–254.
[2] Slaney, M. (1998). Auditory Toolbox, version 2. Technical report, Interval Research Corporation.
[3] Lee, B. S. and Ellis, D. P. W. (2012). Noise Robust Pitch Tracking by Subband Autocorrelation Classification. Based on dissertation, Columbia University.