Speech To Song Classification

Emily Graber

Center for Computer Research in Music and Acoustics, Department of Music, Stanford University

Abstract

The speech to song illusion is a perceptual phenomenon where listeners perceive the transformation of certain speech clips into song after approximately ten consecutive repetitions of the clips. Both perceptual and acoustic features of the audio clips have been studied in previous experiments. Though the perceptual effects are clear, the features driving the illusion are only known to relate to isolated acoustic features. In this paper, speech clips are examined from a music theoretical viewpoint; typical music-theoretic rules are used to derive context dependent features. The performance of classification trees is then used to assess the utility of the music-theoretically-derived features by comparing them to spectral features and linguistic features. Contour features are found to differentiate the speech clips into transforming and non-transforming variants suggesting that music-theoretic schema may be responsible for driving the perceptual classification.

Introduction

The Speech to Song (STS) illusions is a perceptual phenomenon where listeners perceive the transformation of a given *speech* clip into song after approximately ten consecutive repetitions of the clip [Deutsch et al., 2011]. Listeners do not perceive this transformation for all speech clips, thus several perceptual and neuro-imaging studies have aimed to figure out what the perceptual difference *is* between the clips that transform and the clips that do not transform [Deutsch et al., 2011], [Tierney et al., 2013], [Hymers et al., 2015]. These studies were able to find significant differences in behavioral responses and brain responses to transforming and non-transforming (or not-yet-transformed) stimuli.

Given the neural and behavioral difference between transforming and non-transforming stimuli, it is also of interest to know *what* about the stimuli drives the STS illusion. Tierney et al. used statistically matched stimuli for each group (transforming and non-transforming) such that average syllable length, average syllable rate, and average fundamental frequency differences between the groups were not perceptually significant [Tierney et al., 2013]. Within-syllable frequency change and inter-accent intervals were however found to be different between the transforming and non-transforming stimuli though they were not purposely manipulated in the experiment. Margulis et al. explored the relevance of repetition onset timing and semantic/syntactic content for the strength of the STS illusion [Margulis et al., 2015]. As semantics became less and less relevant, the strength of the illusion increased. Falk et al. also found that certain pitch and rhythmic properties facilitated the STS illusion in their careful manipulations of just two clips [Falk et al., 2014]. Most notably, stable within-syllable pitch and perfect fifth jumps made the STS illusion more likely.

It seems that musical features and the ability to access those features drives the STS illusion. In the present study, I explored the naturalistic stimulus set used by Tierney et al.; I differentiated the stimuli based on music-theoretic features, i.e. context dependent features rather than linguistic, semantic, rhythmic, or pitch features alone. Seven feature categories (linguistic, rhythmic, harmonic, contour, pitch, spectral, and general) each with several features were evaluated in terms of their LOOCV test error in classification trees that predicted the perceptual class of the test stimulus, i.e. transforming or non-transforming. Contour features were found to be the best predictors of stimulus type; this supports the notion that context helps drive the STS illusion.

Related Work in Machine Learning

Differentiating speech from music is a common machine learning task. Usually, spectral features like MFCCs, centroid, flux, and tilt, extracted from time domain signals are useful for discriminating between speech and music [Scheirer and Slaney, 1997]. This works because most music contains instrumental contributions which have very different spectral characteristics from the speaking voice. Indeed, spectral features are useful for classifying different musical genres without voice as well. Mandel et al. were even able to classify individual artists by retaining detailed information about full audio clips, i.e modeling unaveraged MFCCs for each clip as a mixture of Gaussians [Mandel and Ellis, 2005]. Nam et al. took an unsupervised learning approach to find useful features for music tagging/annotation/classification [Nam et al., 2012].

In doing so, they were able to use a simple linear classifier to distinguish genres. This method is compelling because the features were not hand crafted as MFCCs and most other spectral features are.

It is challenging to find features that are useful for discriminating between the speaking voice and the singing voice because spectral information is no longer highly informative. Thompson developed a successful method to classify speaking and singing based on pitch stability and pitch probability distributions [Thompson, 2014]. However, in the present application, all audio signals are recorded speech, therefore a different method for feature extraction must be used.

Pitch tracking and onset detection algorithms, used in music information retrieval tasks, are useful for parsing time-domain audio into note-like units. Lee and Ellis developed a robust pitch tracking algorithm for speech that uses a multi-layer perceptron classifier to eliminate octave errors and noise errors that typically plague autocorrelation pitch trackers [Lee and Ellis, 2012]. Lee and Ellis' algorithm also finds the probability that the speech in each time frames is voiced or unvoiced. The start of voiced segments is often analogous to note-onset times. The findings of Falk et al. support this idea as they found that intervocalic interval stability was more important than intersyllabic interval stability [Falk et al., 2014].

Dataset and Features

Stimuli: 48 suitable STS clips with mean duration 1.3859 seconds (SD = 0.3923) were excerpted from audiobook recordings. These clips were previously evaluated in a behavioral and functional imaging study, thus correct labelings were known [Tierney et al., 2013]. Differences between average duration, syllable rate, syllable length, fundamental frequency, phonetic content, and semantic structure were considered and found effectively insignificant between the transforming and non-transforming clips. All clips were mono recordings with 44100 Hz sampling rate.

Processing: All audio processing was done in MATLAB; Lee and Ellis' Subband Autocorrelation Classification (SAcC) was used for initial pitch and onset detection estimates [Lee and Ellis, 2012]. Full transcriptions were made by hand to correct any errors in SAcC, and all features were derived from those transcriptions with the exception of the spectral features. The mean MFCC vectors were obtained by averaging the 13-dimensional MFCCs made from 20 ms Hann windows with 50% overlap calculated by the Auditory Toolbox [Slaney, 1998]. Figure 1 shows an example of estimated pitches, estimated onsets, and a full transcription for a transforming clip.

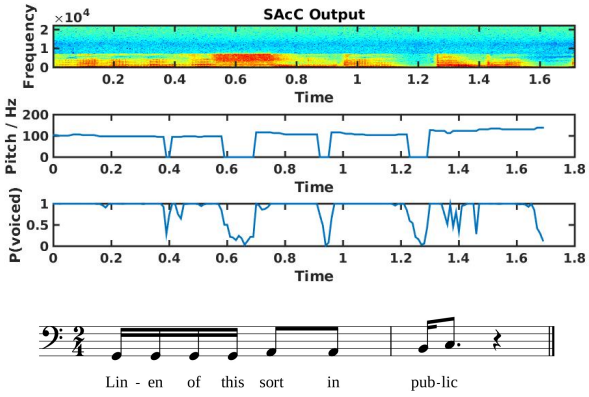

Figure 1: Output of SAcC. From top to bottom: Spectrogram; Pitch estimates; P(voiced) – the probability that the phoneme being spoken is voiced, i.e., vowel-like; Full transcription of the clip – speaker said "Linen of this sort in public."

All feature categories and features are summarized in Table 1 below.

Table 1: Feature Descriptions

Feature Category	Features
Linguistic	number of syllables number of stressed syllables longest word
Rhythmic	total number of onsets number of strong beats pickups syncopations hemiolas implied meter
Harmonic	implied tonic implied dominant implied other mode non-diatonic pitches resolution level resolution strength
Contour	number of melodic leaps number of melodic steps largest leap size in semitones number of consecutive leaps
Pitch	histogram of scale degrees range in semitones melisma
Spectral	mean MFCCs
General	key number of notes number of unique notes total duration

Table 2: Error Statistics

	LOOCV error	Hit Rate	Miss Rate	False Alarm Rate	Correct Rejection Rate	Precision	Recall
Linguistic	0.7292	0.25	0.75	0.7083	0.2917	0.4615	0.2609
Rhythmic	0.5	0.5833	0.4167	0.5833	0.4167	0.5833	0.5
Harmonic	0.3125	0.7917	0.2083	0.4167	0.5833	0.5758	0.6552
Contour	0.125	0.875	0.125	0.125	0.875	0.5	0.875
Pitch	0.2708	0.6667	0.3333	0.2083	0.7917	0.4571	0.7619
Spectral	0.60417	0.4166666667	0.5833333333	0.625	0.375	0.5263	0.4
General	0.4167	0.5833	0.4167	0.4167	0.5833	0.5	0.5833
All	0.16667	0.875	0.125	0.2083	0.7917	0.525	0.8077

Classification Methods

CART[1]*:* Classification and regression trees work by segmenting the feature space of a dataset into discrete bins. A prediction can made according to which discrete bin a test sample's features fall into. In classification trees (as opposed to regression trees) bin boundaries are determined by recursive binary splitting, a greedy procedure where splits are chosen to maximize node purity at the time of the split [James et al., 2013]. For example, given data $x \epsilon \mathbb{R}^{mxn}$, if the cutpoint s were chosen for predictor x_j, there would be two resulting regions: one region containing all samples where $x_j < s$ and one region containing all other samples where $x_j \geq s$. The goal of the classification tree is to choose s and j such that the resulting regions contain samples from only one class (n.b. this is an ideal case). Now that the class labels for those regions are known, any sample that falls into them can be assigned the appropriate label.

With just one split however, it is likely that the resulting regions will not contain single class labels. In this case, the class that is most common in a region becomes the class label for that region. The classification proportions for a region r can then be calculated for each possible class k.

$$\varepsilon_{rk} = \frac{number\ samples\ with\ class\ k\ in\ region\ r}{number\ samples\ in\ region\ r}$$

The Gini Index G measures the 'node' or region impurity over all classes.

$$G_r = \sum_{k=1}^{K} \varepsilon_{rk}(1 - \varepsilon_{rk})$$

Finally, the classification tree aims to create regions by choosing j and s that minimize the Gini Index. If all the samples in a node or region are from the same class (what we want!), $G_r = 0$. The tree continues to make spits until the nodes are pure, or some threshold has been passed. Therefore the number of splits can serve as an indication of how complicated the classification process was. Additionally, splits closer to the root of the tree can be said to be more important that splits near the leaves of the tree.

Classification trees are easy to interpret, i.e. it is clear which feature was chosen for every split, and what the value of the particular feature was to make the best split. I chose to use classification trees precisely for those reasons.

[1]This description is based on *An Introduction to Statistical Learning* by Gareth James et al.

Results

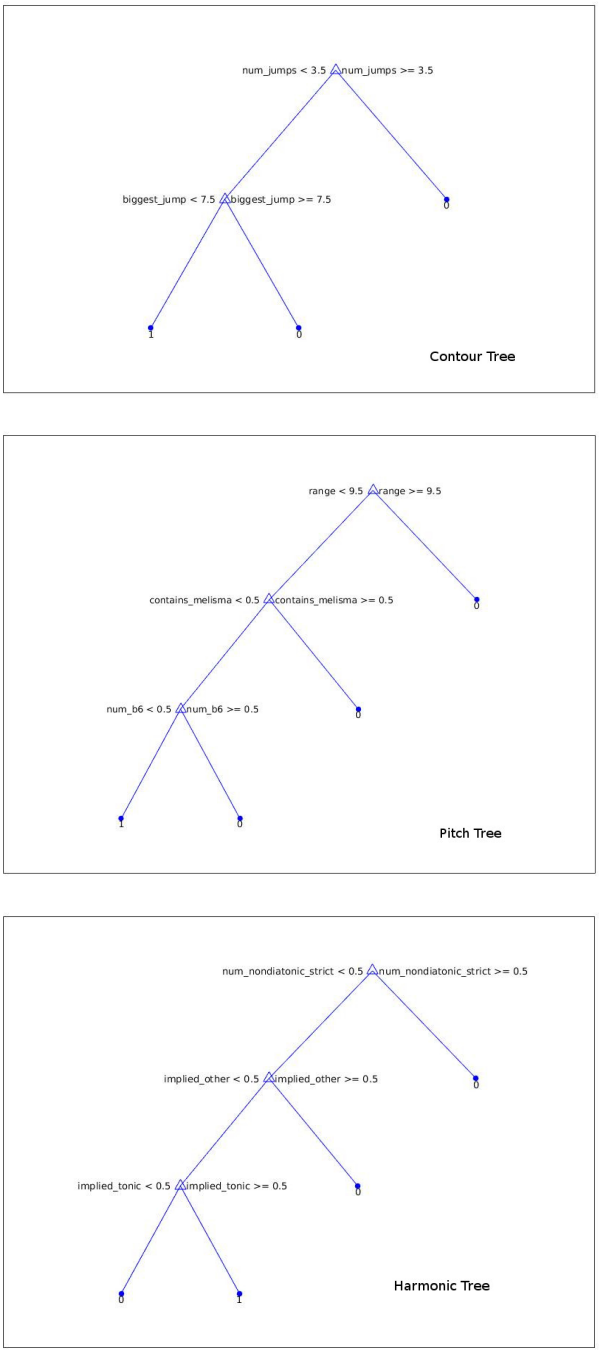

Figure 2: Top to bottom: Classification tree based on Contour Features; Classification tree based on Pitch Features; Classification tree based on Harmonic Features

In order to asses which feature category was most relevant for differentiating the STS stimuli, I created separate classification tress for each category. Because I had a limited set of training data, I choose to evaluate the performance of the trees by leave-one-out cross-validation (LOOCV). Table 2 shows the LOOCV error,

confusion matrix values, and precision and recall metrics. Hits were counted when the test sample turned into song and the prediction was correct. Miss were counted when the test sample turned into song and prediction was incorrect. False alarm were counted when the test sample was not song yet song was predicted (i.e. transforming). Correct rejection was counted when the test sample was not song and the prediction was also not song (i.e. non-transforming).

The three trees with the lowest error and simplest structure are shown in Figure 2. The contour features (number of melodic leaps, number of melodic steps, largest leap size in semitones, and number of consecutive leaps) appear to be the most relevant for differentiating the STS stimuli. The root node divides the stimuli according to the number of jumps that take place in the melody. The second split is based on the largest jump size in the melody. A jump greater than 7.5 semitones (a perfect fifth plus a quarter tone) predicts that the melody will not be perceived as song.

Discussion

The features selected by the trees in Figure 2 support the idea that musical context is playing an important role in the STS illusion. Previous work has shown that pitch stability and jumps of perfect fifths help to improve the STS illusion [Falk et al., 2014]. These features however do not relate to the melody of an STS clip as a whole. A melody is made out of certain pitches with certain rhythms, but the shape of the melody and the tension and release of the melody help to make it sound good or bad, right or wrong. The particular pitches and their placement create the melodic shape and the tension yet they are not identical to shape and tension.

In order to capture the shape of the melody, I created features like 'number of jumps' and 'biggest jump'. To encode the level of tension, I created harmonic features that indicated if the melody contained an implied tonic harmony, dominant harmony, or other harmony because those harmonies index the level of tension and resolution within the melody.

The tree based on contour features shows that the number of jumps within a melody matters. Given that the melodies were under 1.5 seconds, one can imagine that it would be difficult to sing one if it had many large jumps. As Margulis et al. found, it is likely that listeners perceive the illusion more strongly when they can sing along with the melody [Margulis et al., 2015]. The tree based on harmonic features shows that the presence of destabilizing pitches (non-diatonic pitches) is also important in differentiating the transforming and non-transforming clips. These pitches make the underlying key less clear.

More work should be done, but these findings suggest that context is important to the perception of the STS illusion.

Conclusion

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.

Based on the results, the feature set which best classifies the STS stimuli is the melodic contour 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). Therefore oft-repeated music-theoretic schema my help listeners perceive the STS illusion for those stimuli that are music-theoretically well-formed. The role of speaking now needs to be disentangled from the role of context and rule following in this perceptual phenomenon.

References

[Deutsch et al., 2011] Deutsch, D., Henthorn, T., and Lapidis, R. (2011). Illusory transformation from speech to song. *J. Acoust. Soc. Am.*, 129(4):2245–2252.

[Falk et al., 2014] Falk, S., Rathcke, T., and Bella, S. D. (2014). When Speech Sounds Like Music. *Journal of experimental psychology. Human perception and performance*, 40(4):1491–1506.

[Hymers et al., 2015] Hymers, M., Prendergast, G., Liu, C., Schulze, A., Young, M. L., Wastling, S. J., Barker, G. J., and Millman, R. E. (2015). Neural mechanisms underlying song and speech perception can be differentiated using an illusory percept. *NeuroImage*, 108:225–233.

[James et al., 2013] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, New York, 1st edition.

[Lee and Ellis, 2012] Lee, B. S. and Ellis, D. P. W. (2012). *Noise Robust Pitch Tracking by Subband Autocorrelation Classification*. Based on dissertation, Columbia University.

[Mandel and Ellis, 2005] Mandel, M. I. and Ellis, D. P. W. (2005). Song-Level Features and Support Vector Machines for Music Classification. In Reiss, J. D. and Wiggins, G. A., editors, *International Society for Music Information Retrieval conference*, pages 594–599.

[Margulis et al., 2015] Margulis, E. H., Simchy-gross, R., and Black, J. L. (2015). Pronunciation difficulty , temporal regularity , and the speech-to-song illusion. *Frontiers in Psychology: Auditory Cognitive Neuroscience*, 6(Article 48):1–7.

[Nam et al., 2012] Nam, J., Herrera, J., Slaney, M., and Smith, J. (2012). Learning Sparse Feature Representations for Music Annotation and Retrieval. In *International Society for Music Information Retrieval*, number Ismir, pages 565–570.

[Scheirer and Slaney, 1997] Scheirer, E. and Slaney, M. (1997). Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. IEEE.

[Slaney, 1998] Slaney, M. (1998). Auditory Toolbox, version 2. Technical report, Interval Research Corporation.

[Thompson, 2014] Thompson, B. (2014). Discrimination between singing and speech in real-world audio. *MIT Lincoln Laboratory*, pages 407–412.

[Tierney et al., 2013] 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.