Phase Based Features for Singing Voice Detection using Deep Neural Networks

Rupak Vignesh Swaminathan; *School of Music, Georgia Institute of Technology*

Motivation

- Singing Voice Detection is an important subtask in MIR problems such as musical source separation and singing pitch detection.
- MFCCs and other spectral features like spectral centroid, spectral flux, etc. are widely used features for MIR problems including singing voice detection.
- Research questions:
  - Why is this problem challenging? *Audio is Polyphonic*.
  - Is it possible to design features that outperform the standard spectral features for this specific task? *Perhaps*.
- Phase spectrum is often neglected although it contains useful information.

Related Work

- MFCCs, LPCs, standard spectral time domain features such as Spectral Centroid, Zero Crossing Rate, etc as input to Support Vector Machine (SVM) and output probability smoothing with Hidden Markov Models (HMM).
- Aggregation of features over longer window length as vocals are changing rapidly but background is periodic.
- Augmenting the data with label preserving audio transformations (such as pitch shifting, time stretching, adding Gaussian noise, etc).
- The state of the art uses multiple instance learning and saliency maps on the output of a CNN to learn weakly labelled data with raw spectrogram as feature.

Additive Nature of Phase

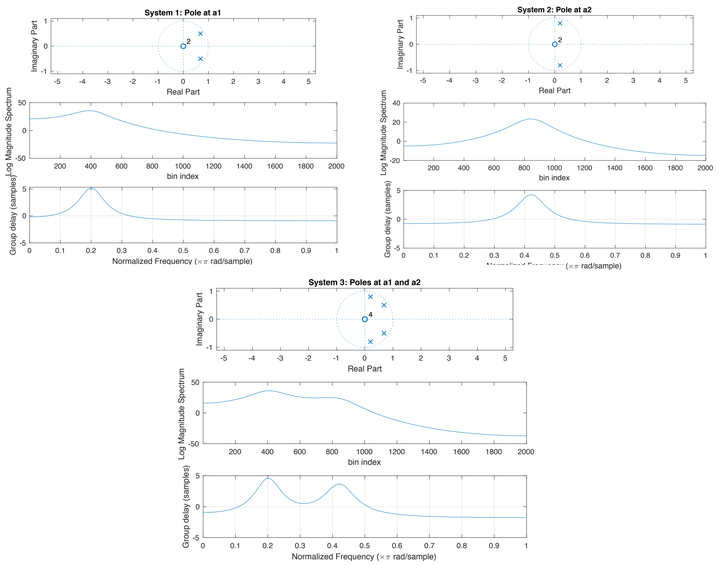

Figure 1. Combinations of two systems is multiplicative in the magnitude spectrum but additive in the phase spectrum

Minimum Phase Group delay

- Group delay can be computed by taking the negative derivative of the unwrapped phase.
- This is very noisy owing to phase unwrapping.
- Also the peaks and valleys of the spectrum will be resolved properly only if the group delay is derived from a minimum phase signal.
- Group delay can also be computed using the differentiation property of the Fourier transform using the following equation.

$$\tau(\omega) = (Xr(\omega)Yr(\omega) + Xi(\omega)Yr(\omega)) \,/\, |S(\omega)|^2$$

- $Xr(\omega)$ and $Xi(\omega)$ are the real and imaginary parts of the FFT of the windowed signal x(n). $Yr(\omega)$ and Yi are the real and imaginary parts of the FFT of windowed signal x(n) multiplied by n.
- $|S(\omega)|^2$ is the power spectrum of x(n).

Phase based features

- The power spectrum in MFCC computation is replaced by the group delay spectrum.
- Other features such as spectral centroid, spectral flux, etc are computed from the group delay spectrum.

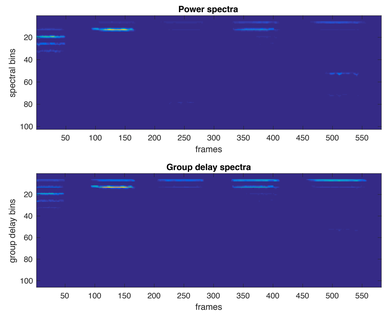

Figure 2. Formants highlighted in the group delay domain (Audio file corresponds to different vowels).

Dataset and Classifier

- The evaluation is done on two datasets, namely the MIR1K and Jamendo datasets.
- The Jamendo dataset contains of 61+16+16 songs (train, valid and test).
- MIR1K dataset contains 1000 snippets which are then manually split into 70% train and 30% test.
- Support Vector Machine (SVM) and Neural networks are used as classifiers for this task.
- The Jamendo contains overlapping vocals and instruments, and is used for testing the algorithm on polyphonic audio.
- The MIR1K data is used for testing on monophonic singing voice.

Results

Table 1.a

	MIR1K (Monophonic)	
	Baseline features (Acc. %)	Proposed features (Acc. %)
SVM C=1	94.77	94.05
NN layers=1	95.6808	95.1193
NN layers=2	95.371	94.8836
NN layers=3	95.387	94.3431

Table 1.b

	Jamendo (Polophonic)	
	Baseline features (Acc. %)	Proposed features (Acc. %)
SVM C=1	71.09	69.05
NN layers=1	75.6333	74.5915
NN layers=2	75.0102	73.5172
NN layers=3	75.0762	72.5452

Table 1.a and 1.b show the test accuracies of MIR1K and Jamendo respectively

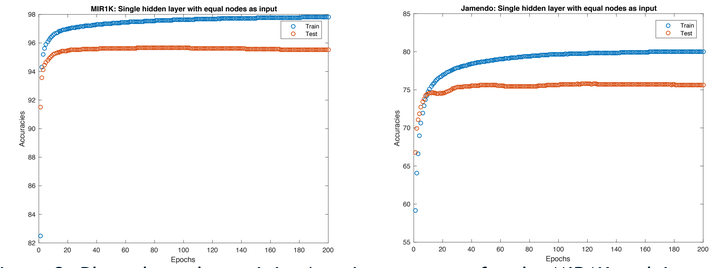

Figure 3. Plots show the training/testing accuracy for the MIR1K and Jamendo dataset respectively.

Discussion and Conclusion

- The test accuracies for both polyphonic and monophonic are not better than the baseline features.
- The MFCCs are usually very robust features and highlighting the formants did not contribute to adding any information.
- In a pilot experiment, the MFCCs were kept intact and the other features such as Spectral Centroid were computed from the group delay spectra. This was not better than the baseline features as well.
- This suggests that the phase based features contain almost the same information as the magnitude spectrum.

Recommendations and Future Work

- Use Neural network as a feature extractor (Spectrogram input, CNNs)
- Aggregate Features over a large time window as vocals change rapidly but the background track is periodic.
- Perform Independent Component Analysis/Matrix Factorization and extract group delay features from spectral templates.

Contact Information

Rupak Vignesh Swaminathan
GeorgiaTech Center for Music Technology
rvs@gatech.edu