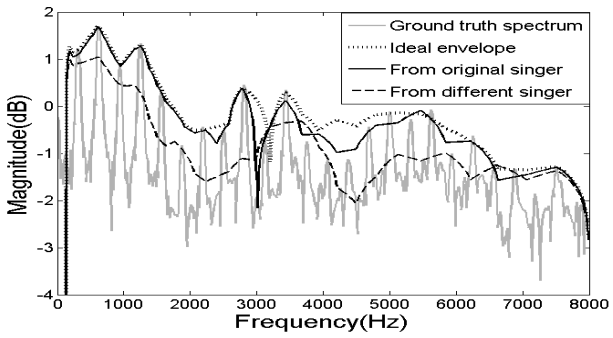


**Figure 3**. Comparison of the reconstructed phoneme envelope of the phoneme /a/ obtained from training mixtures of the same singer and with that obtained from pure vocals of a different singer.

ideal envelope of the phoneme.

Although the NMF optimization converges slowly, the number of iterations to be carried out to obtain the envelope is low, for both training and testing procedures. It is observed that the bases and envelopes attain their final structure after 4000 and 1000 iterations, respectively.

## 7. CONCLUSION

Soft masks derived from a dictionary of singer-vowel spectra are used to improve upon the vocal-instrumental music separation achieved by harmonic sinusoidal modeling for polyphonic music of the particular singer. The main contribution of this work is an NMF based framework that exploits the amply available original polyphonic audios of the singer as training data for learning the dictionary of singer spectral envelopes. Appropriate constraints are introduced in the NMF optimization for training and test contexts. The availability of lyric-aligned audio (and therefore phone labels) helps to improve the homogeneity of the training data and have a better model with fewer basis vectors. Significant improvements in reconstructed signal quality are obtained over binary masking. Further it is demonstrated that a vowel-dependent soft mask obtained from clean data of a different available singer is not as good as the singer-vowel dependent soft mask even if the latter is extracted from polyphonic audio.

## 8. ACKNOWLEDGEMENT

Part of the work was supported by Bharti Centre for Communication in IIT Bombay.

## 9. REFERENCES

[1] L. Boucheron and P. De Leon. On the inversion of mel-frequency cepstral coefficients for speech enhancement applications. In *Int. Conf. Signals Electronic Systems, 2008.*, pages 485–488, 2008.

[2] D. Fitzgerald. Vocal separation using nearest neighbours and median filtering. In *Irish Signals Systems Conf.*, 2012.

[3] J. Han and B. Pardo. Reconstructing completely overlapped notes from musical mixtures. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process., 2011 (ICASSP '11.)*, pages 249 – 252, 2011.

[4] M. Kim, J. Yoo, K. Kang, and S. Choi. Nonnegative matrix partial co-factorization for spectral and temporal drum source separation. *IEEE Journal Selected Topics Signal Process.*, 5(6):1192 – 1204, 2011.

[5] A. Lefevre, F. Bach, and C. Fevotte. Semi-supervised NMF with time-frequency annotations for single-channel source separation. In *Proc. Int. Soc. Music Information Retrieval (ISMIR 2012)*, pages 115–120, 2012.

[6] Y. Li and D. Wang. Separation of singing voice from music accompaniment for monaural recordings. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(4):1475–1487, 2007.

[7] B. Raj, R. Singh, and T. Virtanen. Phoneme-dependent NMF for speech enhancement in monaural mixtures. In *Proc. Interspeech*, pages 1217–1220, 2011.

[8] V. Rao, C. Gupta, and P. Rao. Context-aware features for singing voice detection in polyphonic music. In *Proc. Adaptive Multimedia Retrieval*, Barcelona, Spain, 2011.

[9] V. Rao and P. Rao. Vocal melody extraction in the presence of pitched accompaniment in polyphonic music. *IEEE Trans. Audio, Speech, Lang. Process.*, 18(8):2145–2154, 2010.

[10] M. Ryynanen, T. Virtanen, J. Paulus, and A. Klapuri. Accompaniment separation and karaoke application based on automatic melody transcription. In *IEEE Int. Conf. Multimedia Expo*, pages 1417–1420, 2008.

[11] M. Schmidt and R. Olsson. Single-channel speech separation using sparse non-negative matrix factorization. In *Proc. of Interspeech*, pages 2617–2614, 2006.

[12] J. Sundberg. The science of the singing voice. *Music Perception: An Interdisciplinary Journal*, 7(2):187 – 195, 1989.

[13] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(4):1462–1469, 2006.

[14] T. Virtanen, A. Mesaros, and M. Ryynänen. Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music. In *ICSA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Brisbane, Australia, 2008.

[15] G. Zhou, A. Cichocki, Q. Zhao, and S. Xie. Nonnegative matrix and tensor factorizations: An algorithmic perspective. *IEEE Signal Process. Magazine*, pages 54–65, 2014.