BLIND SOURCE SEPARATION
OF SPEECH AND BACKGROUND MUSIC
FOR IMPROVED SPEECH RECOGNITION

P. Vanroose *

Katholieke Universiteit Leuven, div. ESAT/PSI
Kasteelpark Arenberg 10, B–3001 Heverlee, Belgium
Peter.Vanroose@esat.kuleuven.ac.be

This paper considers the problem of improving the automatic speech recognition of audio fragments containing background music. This problem is put in the framework of linear source separation, where the music component is subtracted from the signal, thereby aiming at a better speech recognition, not necessarily a better subjective audio quality.

PROBLEM FORMULATION

Consider the setup where one wants to apply ASR in the presence of background music, as is often the case with broadcast news or documentary programmes. The speech of these is of good quality, a human has no problem understanding it, but an ASR system (although very robust against moderate white noise) fails terribly when the noise is more structured, since music is disturbing the frequency spectrum in a selective way.

Typical audio fragments of this type can be found at e.g. the news website of the BBC (`http://news.bbc.co.uk/cbbcnews/`) or the news flashes on VRT radio (see `http://www.radio1.be`) or NOS (see `http://omroep.nl/nieuws`), but also any documentary programme will add background music to most, if not all, commentary, since this gives the programme a "juicier" flavour.

The type of music used for this purpose is such that it does not disturb human understanding of the spoken content and does not distract attention. Hence often the music consists of relatively long constant frequency tones, as can be seen on the spectrogram in Figure 1.

In the example of Figure 1, an audio fragment of 31 seconds from a BBC documentary was sampled at 16 kHz and the FFT spectrum was calculated on 512 samples

* This work was partially supported by the K.U.Leuven Research Fund and by the E.U. 5th Framework project "MUSA".

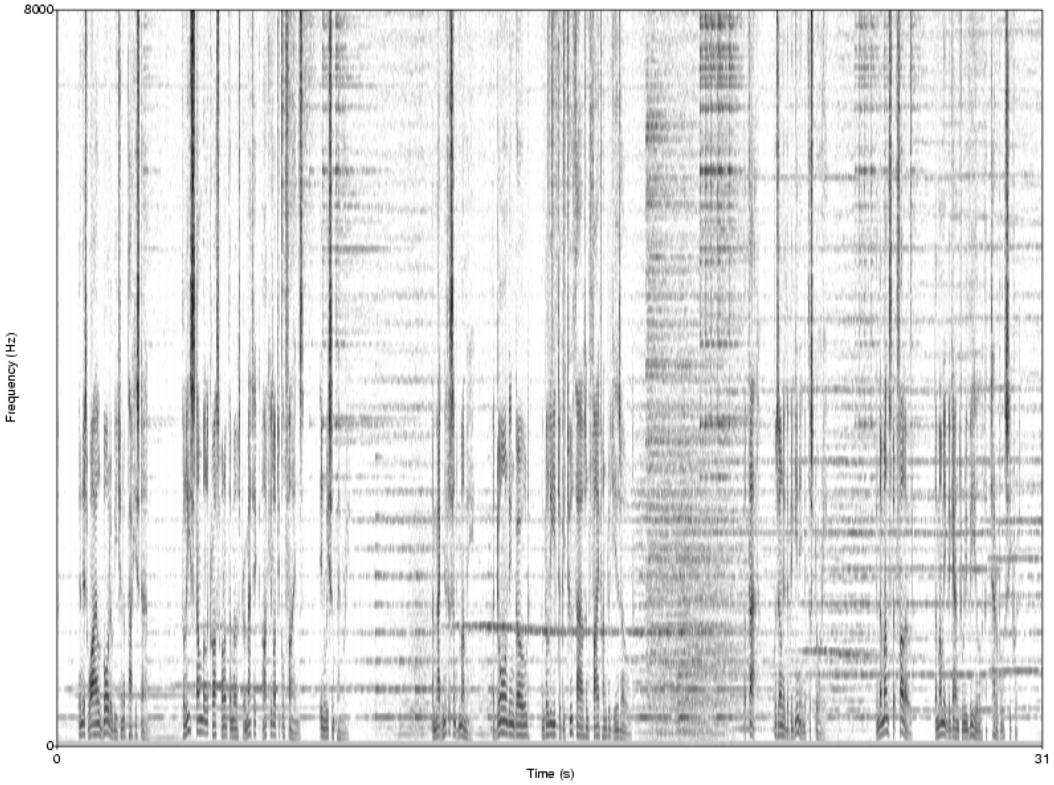

Figure 1: Spectrogram of a 31 second audio fragment with mixed voice and music

at a time, with a time shift of 10 ms. These are standard values for speech recognition, see e.g. [1]. The spectrogram plots these FFT values (vertical axis) against time (horizontal axis), where darker pixels correspond to higher FFT values.

From this spectrogram of a combined voice + music audio fragment, the music contribution, and especially the most disturbing aspect of it for ASR, is visible as horizontal lines.

A speech recogniser will typically start from a simplified version of the spectrogram, viz. the Mel-scale cepstral coefficients, which essentially downsample each column of the spectrogram in a nonlinear way to 24 values, see the upper one third of Figure 2. These 24 values per time instant are here visualised in a similar way as for the spectrogram.

Since not only Mel values but also their temporal changes are important for speech recognition, Figure 2 also visualises the 48 additional feature values often used by a speech recogniser, viz. the first and second order time derivatives of the Mel cepstrum, resulting in 72-dimensional feature vectors, one per 10 ms time frame, which are used to perform a Viterbi search in an HMM phoneme network [2].

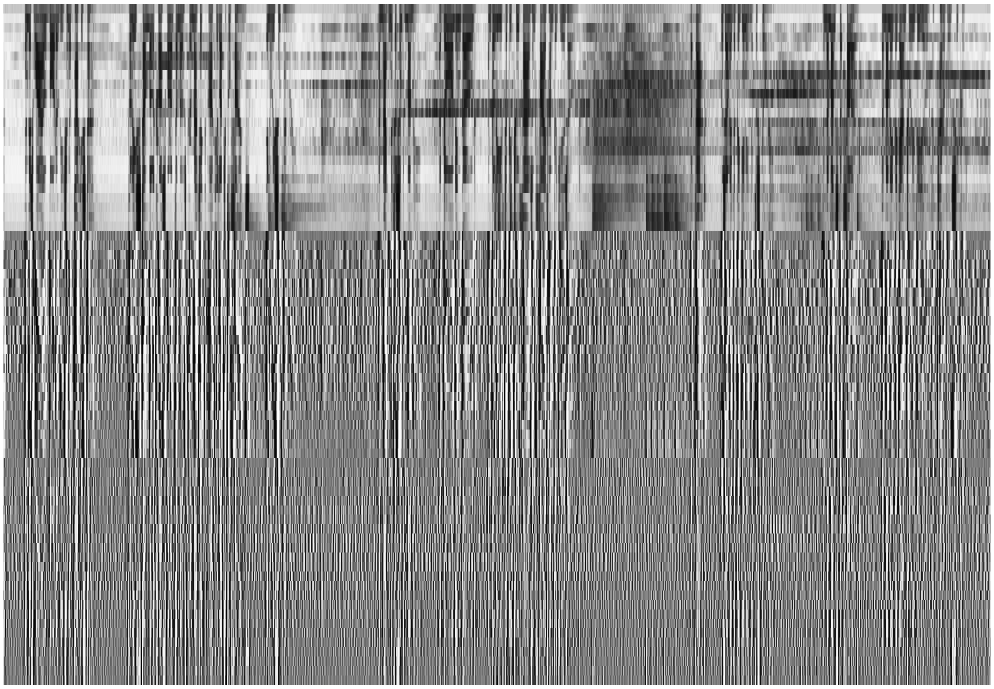

Figure 2: Mel-scale cepstral coefficients and derivatives of the same audio fragment

It is clear that the long horizontal lines in the spectrogram, stemming from the music, are still present in the HMM features, hence it is no surprise that the Viterbi search will map these to completely different phonemes, as opposed to the situation with moderate white Gaussian noise where the Mel cepstral coefficients are much less affected.

A RELATED PROBLEM

Removal of additive noise from a signal is a widely studied problem [3]. It is typically tackled by separating signal and noise, thereby making use of assumed statistical properties of both, like the stationarity of the noise, or the statistical independence of signal and noise, plus of course some discriminative feature which distinguishes the two, e.g. the assumption that the noise is white.

The noise removal problem can be seen as a special case of the more general *blind source separation* problem [4], whereby the sum of two or more statistically independent (stationary) sources is observed and the task is to extract the components.

This differs from the related problem of speech/music *segmentation*, as considered in e.g. [5, 6], where the task is to label audio fragments as either speech or music. That problem can be tackled directly within the ASR HMM framework, by training the "silence" state with music fragments. Unfortunately, a similar approach for background music reduction (applied to the other HMM states) fails, probably because the thus trained phoneme models start to overlap too much.

But from this speech/music segmentation problem we can learn which discriminative features are successful, and use these as guidelines to solve the source separation problem. In [5] two useful features are found to be the momentary *frame entropy* and the *change rate* of the audio. From [6] we learn that music fragments are more "ordered" than speech, i.e., the entropy of music is much more constant in time than the entropy of speech.

INDEPENDENT COMPONENT ANALYSIS

Our task is thus to subtract the music component from the signal, where we may assume that the music and speech components are statistically independent. *Independent component analysis* (ICA) [7] is a well-known technique to do this in an unsupervised way. The efficient FastICA algorithm [8] (for which a freely downloadable matlab version exists) was applied to the 72-dimensional feature vectors. The result is a decomposition of these feature vectors into 72 linear components which are as independent from each other as possible.

The 72 independent components cannot be used directly, since on the one hand we only need two components, and on the other hand ICA does not tell us which component(s) come from the speech part of the signal and which from the music.

Here the results from the speech/music segmentation problem can be used: we assume that music has a more constant frame entropy than speech. Applying ICA to several audio fragments, it was found that the best ASR results were obtained when subtracting from the signal the 10 ICA components with the lowest entropy in the 24 first order derivative components.

The obtained improvement in speech recognition result is not spectacular: the word error rate drops from 45% to around 35%, which is still far to high, but at least this proves that blind source separation of speech and music can be useful for improving speech recognition.

CONCLUSION

Removal of background music from mixed speech/music audio fragments is a difficult problem. A first attempt to use ICA for this purpose shows that it is indeed possible to reduce the speech recognition word error rate substantially.

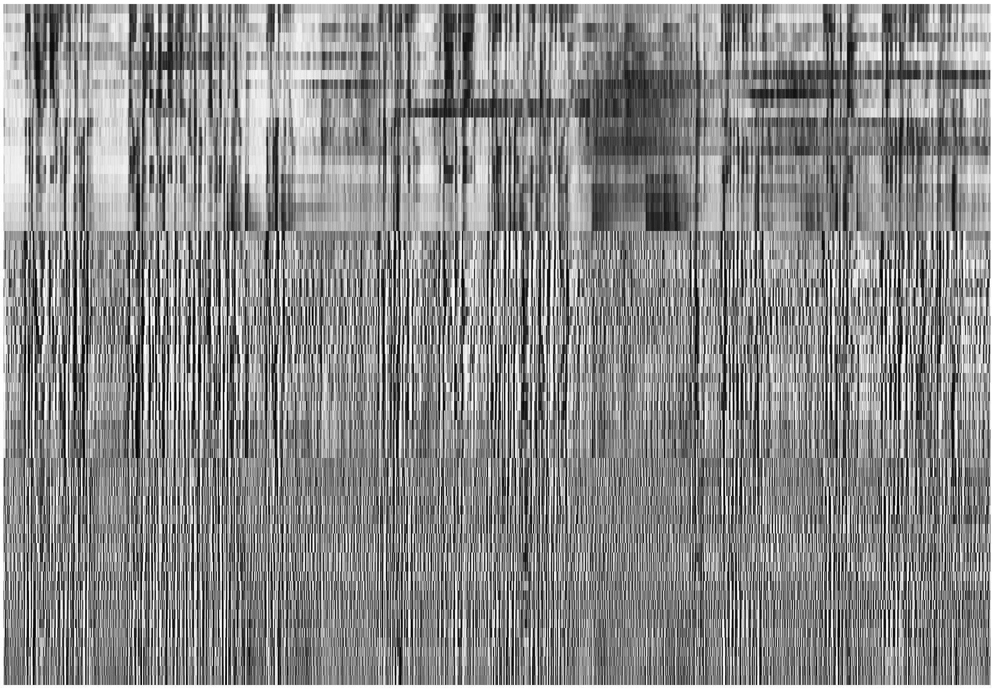

Figure 3: ICA “signal” component of the Mel cepstrum

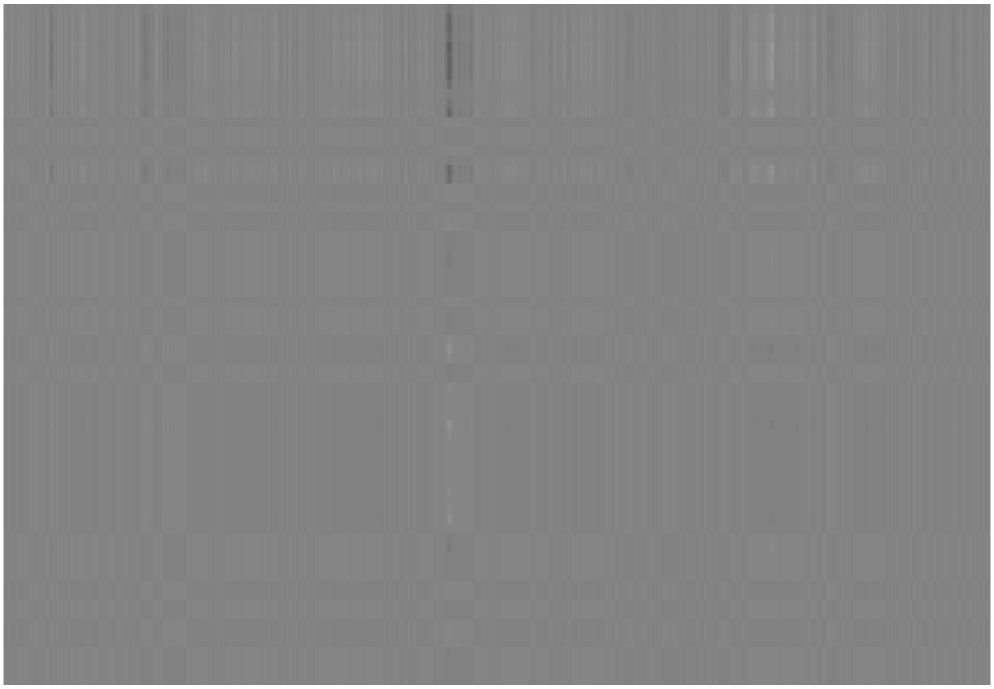

Figure 4: ICA “music” component of the Mel cepstrum

REFERENCES

[1] P. Vanroose, G. Kalberer, P. Wambacq, L. Van Gool, “From speech to 3D face animation”, in *Proceedings of the 23rd Symposium on Information Theory in the Benelux*, Louvain-la-Neuve, pp. 255–260, 2002.

[2] X.D. Huang, Y. Ariki, M.A. Jack, *Hidden Markov Models for speech recognition*, Edinburgh University Press, 1990.

[3] K. Hermus, P. Wambacq and D. Van Compernolle, “Improved noise robustness for speech recognition by adaptive SVD-based filtering”, in *Proc. 20th Symp. on Inform. Theory in the Benelux*, Enschede, pp. 117–124, 1999.

[4] T.W. Lee, M.S. Lewicki, M. Girolami and T.J. Sejnowski, “Blind source separation of more sources than mixtures using overcomplete representations”, *IEEE Signal Processing Letters*, 4(4), pp. 87–90, 1999.

[5] J. Ajmera, I. McCowan, H. Bourlard, “Speech/music segmentation using entropy and dynamism features in a HMM classification framework”, *Speech Communication* 40, pp. 351–363, 2003.

[6] J. Pinquier, J. Rouas, R. André-Obrecht, “Robust speech / music classification in audio documents”, *7th International Conference On Spoken Language Processing* (ICSLP), pp. 2005–2008, 2002.

[7] P. Comon, “Independent component analysis, a new concept ?”, *Signal Processing*, 36(3), pp. 287–314, 1994.

[8] A. Hyvärinen, J. Karhunen, E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.