Letter to the Editor

Image Processing of Speech with Auditory Magnitude Spectrograms

Tilman Horn

Electroacoustics and Audiocommunication, Technical University, München, Germany*

PACS no. 43.72.Ar

Magnitude spectrograms are the standard in speech visualization. In their digital form, they suggest image processing as a powerful tool for speech research. The present article recommends auditory magnitude spectrograms (AMS) for this task, continuing the work of Terhardt [1, 2], Heinbach [3] and Mummert [4, 5]. Three sections give arguments concerning speech representation, interpretation and modification.

1. Representation

In contrast to traditional magnitude spectrograms of constant analysis bandwidth, a straightforward analysis-resynthesis pair can be given for AMS images. The resynthesis method presented below obtains 16bit sound-files of very high quality from 8bit image-files, without restrictions to voiced speech or similar, and without iterative processes [6]. Informal listening tests over headphones yielded faint or no degradation from the CD-quality originals.

1.1. Analysis

Signal-to-AMS conversion is accomplished with the auditory Fourier-t-transformation [2, 3] implemented as a filter bank, each channel consisting of a complex demodulator followed by a low-pass filter. The actually used filter parameters [4, 5] are: quadruple real pole, normalized frequency function, compensated time delay, mathematical 3 dB bandwidth of 0.3 Bark. Only magnitude is evaluated, quantized in 0.5 dB steps. For the resynthesis method presented below, an AMS must have the sampling rate of the time signal and a channel spacing of 0.05 Bark. Convenient storage and image display may require downsampled versions.

Figure 1 shows the lower frequency section of such an AMS.

1.2. Resynthesis

AMS-to-signal conversion is accomplished with the following formula:

$$\hat{s}_i = \sum_{k=1}^{K} \left| S_{ik} \right| \sin \hat{\phi}_{ik}, \tag{1}$$

$$\hat{\phi}_{ik} = \hat{\phi}_{i-1,k+n} + \omega_k T_s, \tag{2}$$

with n such that

$$\left| S_{i-1,k+n} \right| = \max \left\{ \left| S_{i-1,k-N} \right|, \ldots, \left| S_{i-1,k+N} \right| \right\}. \tag{3}$$

To the speech time-signal $\hat{s}_i$, each channel k contributes a sinusoidal signal of the given instantaneous amplitude $|S_{ik}|$. The instantaneous phase $\hat{\phi}_{ik}$ is composed of a predecessor phase and a phase increment. The predecessor phase $\hat{\phi}_{i-1,k+n}$ is taken from the neighbour channel $k+n$ with maximum amplitude $|S_{i-1,k+n}|$; optimal searching width is $N = 5$, closely below the minimal distance of spectral peaks. The phase increment is given by the (angular) channel frequency ω_k times the sampling interval T_s. Phase initialization and handling of multiple amplitude maxima do not play a crucial role.

1.3. Underlying principle

The presented resynthesis method provides spectral phase locking and temporal phase continuity for the micro "hills" of a magnitude spectrogram, so that a reanalysis of a resynthesized spectrogram shows the same hills again. Neither original absolute phases nor original phase relations among the hills are reproduced, but this is not required for auditorily correct resynthesis – as long as the hill ridges (contours, Figure 2) keep certain minimal distances in frequency [1, 3] and time [4, 5]. This can only be ensured by employing an auditory filter bank.

It should be noticed that the presented resynthesis method is not equivalent to resynthesizing just the frequency contours or part tones [3] (horizontal lines in Figure 2). Disregarding the time contours or part clicks [4] (vertical lines in Figure 2) causes tonalization. For direct resynthesis of frequency and time contours *together*, good phase specification has turned out to be rather complicated [5], whereas AMS resynthesis can exploit the inevitable uncertainty of linear analysis as a margin for phase specification, thus allowing the given simple formulation.

Why the small analysis bandwidth of 0.3 Bark? The resynthesis still mishandles the time contours, shifting their energy onto the frequency contours. Thus an analysis with a bias toward the frequency-contour side yields better over-all results, excellent in the case of speech.

2. Interpretation

Purposeful image processing requires image interpretation. This section argues for the expressiveness of AMS images.

2.1. Physiological interpretation

An AMS can be interpreted as a simplified representation of the excitation level along the basilar membrane in the inner ear [7]. Hence features of the neurophysiological periphery can be added to this model, such as aspects of masking (see section 3.1.).

2.2. Psychoacoustical interpretation

Resynthesis of selected AMS parts can demonstrate that the following elementary categories of auditory sensation [7, 8] are associated with characteristic AMS features, reflecting elements of acoustic speech production:

- Harmonicity and virtual pitch are associated with a regular spectral configuration of hills (F0 harmonics), reflecting voiced excitation.
- Noise sensation is associated with an irregular cluster of hills, reflecting fricative excitation.
- Roughness is associated with a regular temporal configuration of hills, reflecting intermittent excitation.
- Timbre is associated with a spectral configuration of prominent hills (formants), reflecting cavity shape.
- Loudness is associated with the integral amplitude of a hill configuration, reflecting source energy.
- Event sensation, subjective duration and rhythm are associated with a temporal configuration of prominent edges, reflecting abrupt changes in source energy.

Received 15 March 1996,
accepted 23 April 1997.

* Contact address: Centrum für Informations- und Sprachverarbeitung, Ludwig-Maximilians-Universität, Oettingenstr. 67, D-80538 München. E-mail: horn@cis.uni-muenchen.de

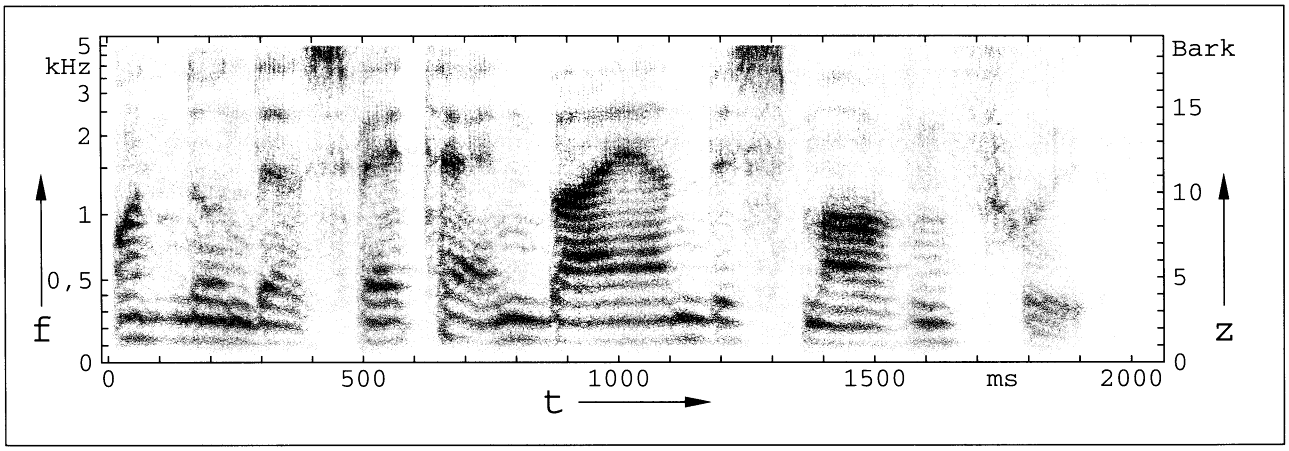

Figure 1. Auditory magnitude spectrogram (blackening indicates amplitude), visualizing an utterance from a German male newscaster: "Und nun das Wetter in Bayern bis morgen früh". An AMS can be easily resynthesized to a speech signal of very high quality (see section 1.2.).

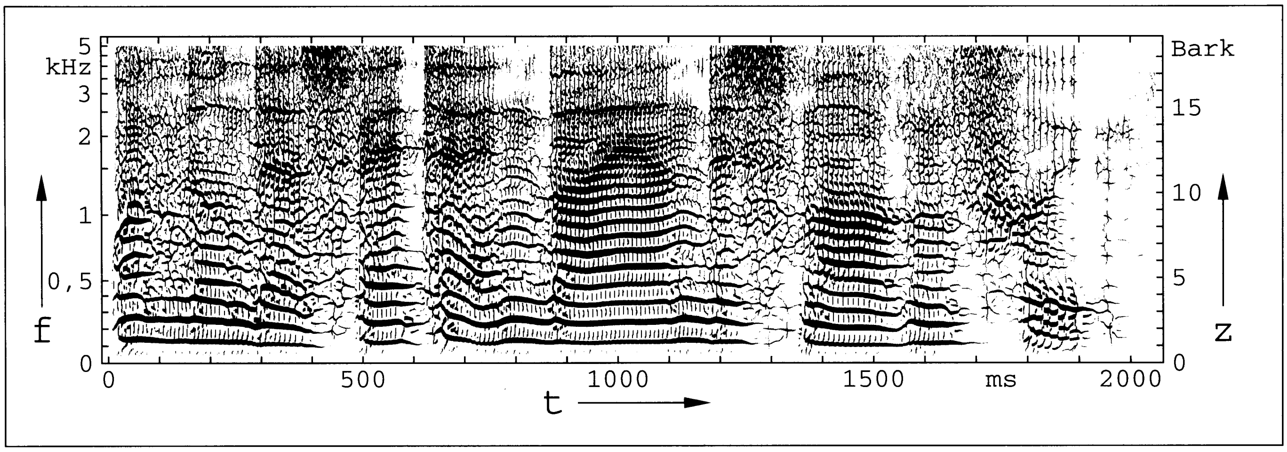

Figure 2. As Figure 1, but contourized to part tones and part clicks (thickness indicates amplitude), representing primary auditory gestalts (see section 2.3.).

The first three AMS features constitute the micro structure, the last three the macro structure, reflecting the excitation-filter model of acoustic speech production (source energy is modelled by filter gain). "Micro" and "macro" should not be taken too literally: the respective ranges of the time-frequency scales overlap to a great extent, and thus the structures cannot be properly separated by simple filtering. However, on the premises of linear non-adaptive analysis, an AMS represents *both* structures quite optimally: the special increasing analysis bandwidth is suited, for example, to both natural movement of harmonics and natural bandwidth of formants, generally outperforming the constant analysis bandwidth – narrowband *and* wideband.

2.3. Gestalt-psychological interpretation

Contourization plays an elementary role as a link between the worlds of physics and information in both visual and auditory perception. The contours in Figure 2 represent primary auditory gestalts, grouping to secondary gestalts, and so on [9].

This motivates the most ambitious interpretation of an AMS as an image which can stimulate processes of higher visual perception *analogous* to processes of higher auditory perception. So via an AMS, methods of cognitive image processing become applicable to speech (see section 3.2.).

3. Modification

Modification and subsequent resynthesis of AMS images yields speech without annoying artifacts, provided that the minimal contour distances are obeyed and no channel signals with overly steep slopes or steps are produced. However, the latter can be repaired by employing a suitable synthesis filter bank [10].

Two basic examples of the use of the above AMS interpretations follow.

3.1. Detail reduction

A simple and quantitatively exaggerated model of spectral and temporal masking can reduce AMS images to a few perceptually most relevant islands (Figure 3). An auditory amplitude threshold in the form of a smoothed AMS is applied, using the above given auditory analysis kernel again, with a multiple of the spread in frequency and time. Moreover, keeping only frequency contours and reducing their time, frequency and amplitude resolution [3], the data rate goes down to a few kbit/s, while resynthesis still produces intelligible speech.

The method suffers from musical noise due to part-tone singularization, but it is more robust than vocoder parametrization of comparable rate and effort. Hence further investigations should be worthwhile.

For instance, the part-tone amplitudes of such a reduced pattern carry much less information than the part-tone frequencies – this could be exploited in hearing aids for recruitment patients, fixing the

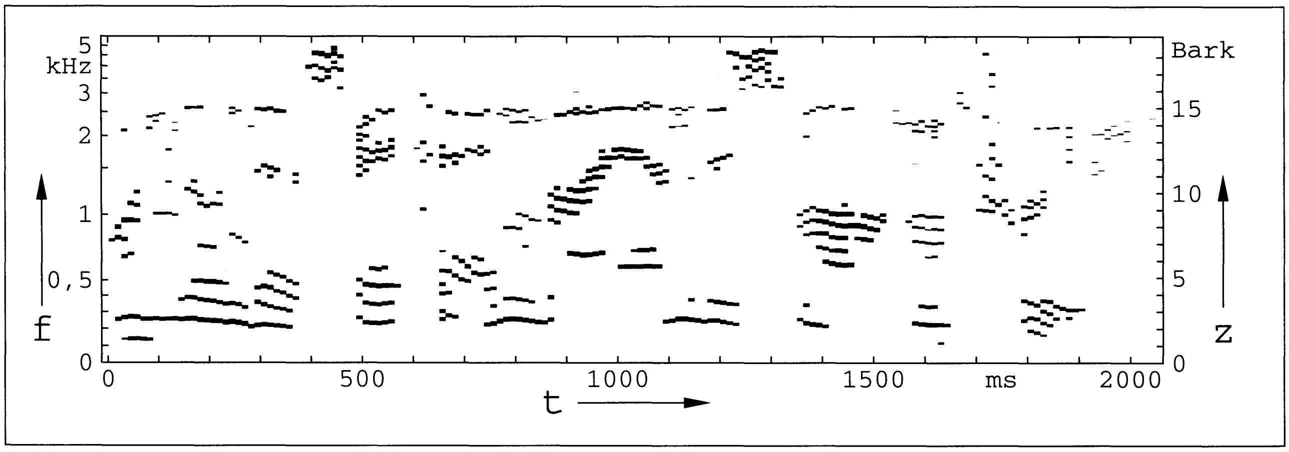

Figure 3. As Figure 2, but reduced to a few elementary part tones, still resynthesizable to intelligible speech (see section 3.1.).

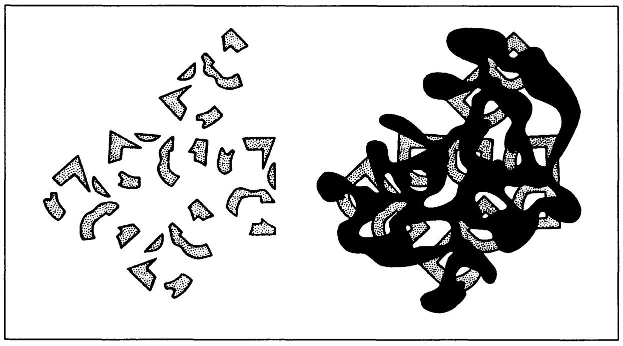

Figure 4. (From Bregman). A visual analogon to the auditory effect of noise clipping: cutting out the black mask is harmful instead of helpful for the recognition (see section 3.2.).

amplitudes to (individually frequency-dependent) values equidistant from the hearing and pain thresholds.

3.2. Noise reduction

The above described amplitude threshold recalls techniques often tried in speech enhancement, aiming to increase the SNR and eventually the intelligibility. Corresponding informal experiments with AMS yield disappointing results: setting to zero the islands dominated by background noise leads to *impaired* intelligibility, as has already been formally demonstrated for a constant-bandwidth representation [11].

It has been concluded that along with the noise, *not completely* masked speech parts contributing to intelligibility are also removed [11]. However, considering the analogy between visual and auditory perception described above, impaired intelligibility can already be explained supposing that nothing audible but noise is removed.

Figure 4 (from Bregman [12]) may be understood as a visual analogon to the auditory effect of noise clipping: cutting out the black mask produces wrong explicit contours and prevents right implicit contours – the "B" symbols cannot be recognized from the processed version (left), but can be from the unprocessed (right). Hence even the removal of *completely* masked parts is harmful instead of helpful for the recognition.

Unfortunately, in demonstrations of noise reduction this effect is often obscured by playing the unprocessed version first ("priming").

References

[1] E. Terhardt: Über die durch amplitudenmodulierte Sinustöne hervorgerufene Hörwahrnehmung. Acustica **20** (1968) 210–214.

[2] E. Terhardt: Fourier transformation of time signals: conceptual revision. Acustica **57** (1985) 242–256.

[3] W. Heinbach: Aurally adequate signal representation: the part-tone-time-pattern. Acustica **67** (1988) 113–121.

[4] M. Mummert: Trennung von tonalen und geräuschhaften Anteilen im Sprachsignal. Fortschritte der Akustik, DAGA '90, DPG-GmbH, Bad Honnef, 1990. 1047–1050.

[5] M. Mummert: Sprachcodierung durch Konturierung eines gehörangepaßten Spektrogramms und ihre Anwendung zur Datenreduktion. Dissertation. Submitted to Fakultät für Elektrotechnik und Informationstechnik, Technische Universität München, 1997.

[6] D. W. Griffin, J. S. Lim: Signal estimation from modified short-time fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing **ASSP-32** (1984) 236–243.

[7] E. Zwicker: Psychoakustik. Hochschultext, Springer-Verlag, Berlin, 1982, (E. Zwicker, H. Fastl: Psychoacoustics – Facts and Models. Springer-Verlag, Berlin, 1990.).

[8] E. Terhardt: Sprachparameter in der Hörwahrnehmung. – In: Interaktion zwischen Artikulation und akustischer Perzeption. M. Spreng (ed.). Georg Thieme Verlag, Stuttgart, 1980, 79–86.

[9] E. Terhardt: From speech to language: on auditory information processing. – In: The Auditory Processing of Speech: From Sounds to Words. M. E. H. Schouten (ed.). Mouton de Gruyter, Berlin, 1992, 363–380.

[10] M. Mummert: Rücktransformation des Kurzzeitspektrums der Fourier-t-Transformation und Ansatz für eine gehörgerechte Transformationskodierung. Fortschritte der Akustik, DAGA '91, DPG-GmbH, Bad Honnef, 1991. 753–756.

[11] J. M. Kates: Speech enhancement based on a sinusoidal model. Journal of Speech and Hearing Research **37** (1994) 449–464.

[12] A. S. Bregman: Auditory scene analysis: the perceptual organization of sound. The MIT Press, Cambridge, Massachusetts, 1990, 25–29.