A COMPARISON OF SPECTRO-TEMPORAL REPRESENTATIONS OF AUDIO SIGNALS

J.D. Krijnders and P.W.J. van Hengel

Cognitive Systems Group, INCAS^3, Assen, the Netherlands

ABSTRACT

This article compares methods for the conversion of time-series into a spectro-temporal representation. These methods are designed based on a resemblance with the auditory processing of sound in the inner ear, or on mathematical principles related to, for example, Fourier analysis. This study provides a comparison between several of these methods.

Two tests were devised for this comparison: one based on susceptibility to noise and one on the expression of spectro-temporal detail. While some methods produced good results on one of the two tests, others produced good results on both.

Overall the transmission line model using an impedance function suggested by Zweig [1] provided the best results, though not significantly. Also a larger computational load may hinder application in some domains. The gammatone filterbank and straightforward spectrogram provide good alternatives with less computational load.

Index Terms— Spectral analysis, Spectrogram, Acoustic signal processing, Signal mapping, Acoustic signal detection

1. INTRODUCTION

Time-frequency processing is the basis of most sound recognition systems and research. Due to its light computational load the short-term Fourier transform is used frequently for this propose. In computational auditory scene analysis, based on the motivation to be more perceptually correct, models the human cochlea are frequently used. In this paper we compare several mathematically motivated methods to several cochlea-model based time-frequency representations. For an overview of the methods compared see table 1.

A comparison could be made on many properties, but in this work we present two properties that are relevant for recognition systems: their spectro-temporal resolution and their susceptibility to noise. The smaller the influence of background noise on the representation of a target signal, the more likely the signal can be properly detected and recognized in noisy real-life conditions. And the more signal-related spectro-temporal detail a representation contains, the more information can be extracted from it, which, in turn, should also increase the chances of the signal being properly detected and recognized. Since no use was made of specific properties of speech signals, e.g. the harmonic or temporal structure of speech, the value of this study for improvement of speech or speaker recognition systems will be limited. As stated before, the focus of this study is on use of conversion methods for more general CASA systems.

INCAS^3 is co-financed by the European Union (European Fund for Regional Development), the Dutch Ministry of Economic affairs (Peaks in the Delta), the Province of Drenthe and the Municipality of Assen.

1.1. Relation to state of the art

The usual approach of comparing recognition results (e.g. [2, 3] on some standard database implies implementing a recognition system on top of the spectro-temporal conversion method. This introduces the possibility that performance is influenced by the interaction between the conversion and the recognition system. One recognition system may be better at using aspects of a certain representation than another. Also, the analysis may be biased toward speech, music or other specific types of sound. To avoid influencing of biasing the results, the performance measures used in this paper are based purely on the output of the conversion method.

To test noise robustness we follow a well-known test for investigating susceptibility of spectro-temporal representations of sound to the effects of noise: the AURORA II test [3]. Although developed for speech recognition and using speech signals as the target exclusively, this test is sufficiently rich in both targets and background sounds to yield indicative results for other sound classes. This test database is thoroughly documented, well calibrated and widely used.

The time-frequency resolution of the transmission line model used in this study was analyzed by Hut et al [4] and compared to that of a gammatone filterbank similar to the one used in this study. The approach taken in that paper was based on the impulse response and its Fourier transform, from which σ_t and σ_ω were computed as defined by Gabor [5]. The outcome $\sigma_t\sigma_\omega$ can be interpreted as an area in the spectro-temporal domain occupied by a simple signal such as a pulse, a tone or tone burst. This area can theoretically never be smaller than $1/2$ and Gabor functions, used in wavelet analysis, produce the minimal area of $1/2$. Computing σ_t and σ_ω and comparing the product of these two values with the minimum value of $1/2$ thus provides a straightforward way of comparing different methods on the resolution of spectro-temporal detail.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Table 1. Overview and subdivision of the techniques used

Name	Abbreviation	# bins	References and/or implementation	Formulae
Short Term FFT	SFFT	512	Octave	$X(\tau,\omega)=\int_{-\infty}^{\infty}x(t)w(t-\tau)e^{-i\omega t}dt$
Mel-Frequency Scaled SFFT	MF-SFFT	100	[6]	
Wavelet	WT	128	[7] own	$G(t,\tau,f_c)=\frac{1}{\sqrt{2\pi}\frac{1}{f_c}}sin(2\pi f_c(t-\tau))e^{\frac{-(t-\tau)^2}{2\frac{1}{f_c}^2}}$
Gammatone filterbank	GT	93	[8] [9]	$g(t,\omega)=at^{\gamma-1}\exp -btcos\omega t$
Linear transmission-line model	LTL	600	[10] own	
LTL With Zweig impedance	LTLZ	600	[11] own	

2. METHODS

2.1. Implementation of conversion methods

Most models were implemented in Octave, only the transmission-line model used a C-implementation based on the original Fortran code by Duifhuis et al [10] and reimplemented by our institute. The number of filters or sections and other settings in each of the methods were the defaults with the implementations provided or based on references. The outcome of the transformation was converted to an energy by the Hilbert transform (WL, GT, LTL, LTLZ) or the absolute value (SFFT, MF-SFFT). All energies were converted to dB.

2.2. Tests of performance

2.2.1. Noise robustness

The noise robustness test is based on the AURORA II test. Here TI-Digits strings are mixed with various types of noise at signal-to-noise ratios from -5 dB to 20 dB. Recordings from a large number of different speakers are available, together with recordings of six different noise types. To keep computational effort within reasonable limits, we decided to use recordings of two male and two female speakers for all digits and mix these with all six noise types. We extended the SNR range used in AURORA II to -30 to +30 in 5 dB steps.

Spectro-temporal representations of energy were computed for the target signal (without noise) $E_t(t,f)$, for the noise signal (without target) $E_n(t,f)$ and for the mixture target+noise $E_m(t,f)$. Some regions of the spectro-temporal domain are dominated by the target signal, whereas other regions are dominated by the noise. The regions dominated by either the target or the noise are found using the following masks:

$$MT(t,f) = (|E_m(t,f)-E_t(t,f)| \le |E_m(t,f)-E_n(t,f)|) \;\&(E_m(t,f)-E_t(t,f)\le 3) \quad (1)$$

$$MN(t,f) = (|E_m(t,f)-E_n(t,f)| < |E_m(t,f)-E_t(t,f)|) \;\&(E_m(t,f)-E_n(t,f)\le 3) \quad (2)$$

We then computed the fraction of the total energy of the target signal that could be traced in the mixed energy: E_t^t, and similarly E_n^t for the noise signal. (N.B. these two fractions need not add up to 1. If the noise and target do not overlap in the time-frequency domain, both values approach one). Figure 1 provides examples of the masks $MT(t,f)$ and $MN(t,f)$ for the gammatone filterbank.

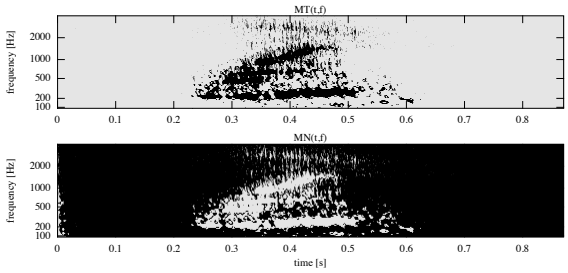

Fig. 1. Masks as specified in equations 1(top) and 2(bottom), respectively the regions dominated by target and noise for MAE_1A target with car noise mixed at 0 dB SNR. The gammatone was used as representation. Black represents ones in the mask while white represents zeros.

Figure 2 shows the values of the fractions E_t^t and E_n^t, computed with the gammatone filterbank for different values

of the SNR for target signal *MAE_1A* and noise signal *car*. The value of E_t^t is 1 for high SNR values, as one would expect. At these values the signal energy should be completely traceable in the representation of the mixed signal. At the lowest SNR values the signal can no longer be traced and is completely submerged in the noise, so E_t^t drops to 0. For E_n^t the situation is almost reversed, approaching 1 at the lowest SNR values. This quantity, however, does not fall to 0 at the other end of the scale. Since the target is concentrated in time and frequency, whereas the noise is more distributed over the time-frequency plane, there will always be parts of the spectro-temporal domain in which the noise will dominate, even at the highest SNR values. This implies that E_n^t does not fall to 0.

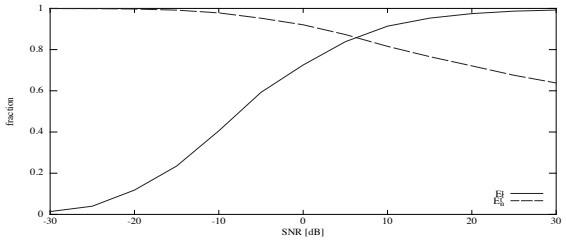

Fig. 2. Fractions E_t^t and E_n^t, computed with the gammatone filterbank for different values of the SNR for target signal *MAE_1A* and noise signal *car*

To allow a direct comparison between different methods on susceptibility to noise, the curve shown in figure 2 can be reduced to a single number by taking the SNR value where E_t^t reaches to 2/3. We will refer to this value as SNR_{thresh} as it is indicative of the threshold above which the amount of traceable energy should be sufficient to allow recognition. The different spectro-temporal distributions of energy of the different noise types lead to sizable differences in the SNR_{thresh} values of these "broadband" noises. We therefore decided not to average the results over the noise type.

2.2.2. Spectro-temporal detail

We adapted Hut's approach and computed estimates for σ_t and σ_ω from the spectro-temporal energy representations of impulse and tones respectively.

For the computation of σ_t we used the spectro-temporal energy representation resulting from stimulation by a single pulse $E_p(t, f)$, and

$$\sigma_t^2(f) = \frac{1}{E(f)^2}\int (t - t_0(f))^2\, 10^{2E_p(t,f)/10} dt \quad (3)$$

$$t_0(f) = \frac{1}{E(f)^2}\int t\, 10^{2E_p(t,f)/10} dt \quad (4)$$

$$E(f)^2 = \int 10^{2E_p(t,f)/10} dt \quad (5)$$

computed for each frequency channel in $E_p(t, f)$. Note that f is stimulus frequency and as the integration is over $-\infty$ to ∞ the factor 2π makes no difference. The power of ten terms convert the decibels back to the linear domain. For the computation of σ_ω we used the spectro-temporal energy representation $E_\omega(t, f)$ resulting from stimulation by sinusoids at frequencies corresponding to the center frequencies of all frequency channels. Thus the number of stimulations is equal the number of bins as specified in table 1. This replaces the Fourier transform of the impulse response in the paper by Hut et. al. ([4]). In this case

$$\sigma_\omega^2(f) = \frac{1}{E(\omega)^2}\int (\omega - \omega_0(f))^2 10^{2E_\omega(t,f)/10} d\omega \quad (6)$$

$$\omega_0(f) = \frac{1}{E(\omega)^2}\int \omega 10^{2E_\omega(t,f)/10} d\omega \quad (7)$$

$$E(\omega)^2 = \int 10^{2E_\omega(t,f)/10} d\omega \quad (8)$$

If we use the gammatone filterbank and the wavelet analysis to compute the energy representations, the product $\sigma_t\sigma_\omega$, shown in figures 6 and 5 resp., approaches the theoretical minimum of 0.5 for the center part of the frequency range as expected. At the edges of the frequency range of the models we see deviations caused by the finiteness of the integration domain. For frequencies near these edges $E(\omega)$ can be regarded as truncated to the integration domain producing deviations in the value of σ_ω. The wavelet analysis can be seen to approach the theoretical minimum more closely than the gammatone filterbank, as expected.

3. RESULTS

3.1. Noise robustness

Figures 3 and 4 show the SNR thresholds as defined in section 2.2.1 for all methods tested. It can be observed that the spectrogram, gammatone filterbank and the linear transmission line with Zweig impedance function produce comparable results, all with thresholds around -5 dB. The linear transmission line produces values around -2 dB. The wavelet scores worst with a values around +2 dB. Noise types *car*, *restaurant* and *train* generally have a higher masking effect, leading to higher SNR thresholds, whereas street noise for all methods results in the lowest thresholds.

3.2. Spectro-temporal detail

Figures 5 and 6 show the curves for the product $\sigma_t\sigma_\omega$ for all different methods. Most curves show the effects of the finiteness of the frequency range as described in section 2.2.2, especially at the high frequency end. Focusing attention on the midfrequency region therefore, where the results are most reliable, most of the methods are grouped around the theoretical

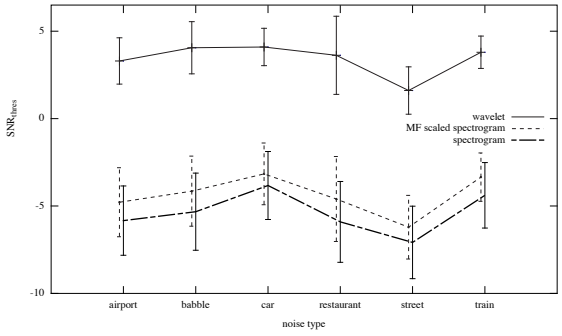

Fig. 3. SNR Thresholds for the mathematically-inspired methods

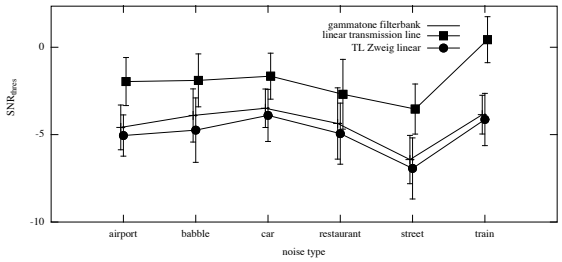

Fig. 4. SNR Thresholds for the biologically-inspired methods

minimum value of 0.5, with the mel-frequency-scaled spectrogram performing slightly worse with values around 1. The obvious exception is the original transmission line model with significantly higher $\sigma_t\sigma_\omega$ values. The linear version with the Zweig impedance performs better and falls in the range of the other methods.The "humps" in the curve of the linear transmission line with Zweig impedance seen around 2 kHz are caused by an "after ringing" of the impulse response.

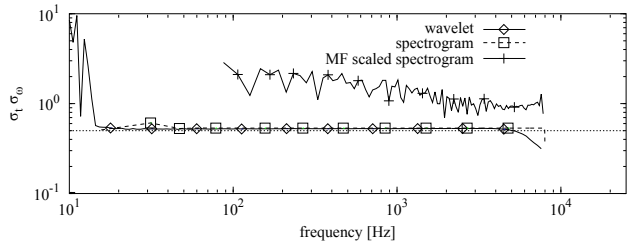

Fig. 5. Curves of $\sigma_t\sigma_\omega$ for the mathematically-inspired methods

4. DISCUSSION

Both the comparison of susceptibility to noise and the comparison of spectro-temporal detail show significant differences between different methods for conversion of waveforms to a spectro-temporal representation of energy.

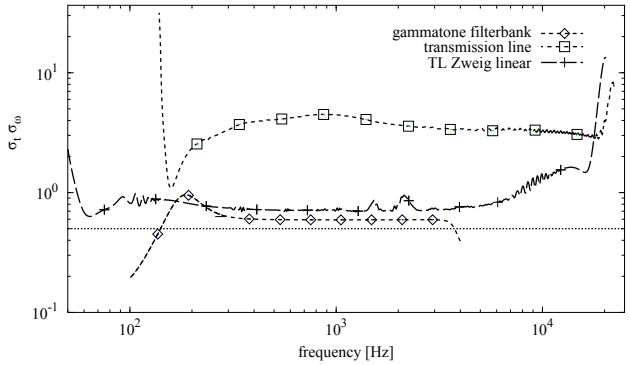

Fig. 6. Curves of $\sigma_t\sigma_\omega$ for the biologically-inspired methods

Similarly, a good performance of the wavelet analysis on the spectro-temporal detail is countered by the worst performance on susceptibility to noise. The gammatone filterbank and spectrogram and the linear transmission line model using the Zweig impedance perform reasonably well on both tests. As could be expected the Mel-frequency scaling of the spectrogram reduces the spectro-temporal detail.

The fact that some methods perform worse than others on the tests presented here does not imply that their use should be avoided. The tests used here are meant to compare different methods for human-centered sound analysis. The aspects of the spectro-temporal representation which are important in a human-centered system depend on the task. However, for a system functioning in a real world environment without prior knowledge about target or noise signals, noise robustness and spectro-temporal detail will most likely be essential. The results on the tests shown here indicate which methods perform better than others on these two aspects. They can not be directly related to human performance on e.g. speech recognition. Although the database used for the test on noise susceptibility uses speech samples and is derived from the AURORA II test for noise-robust speech recognition systems, a direct comparison with speech recognition scores should be avoided. Recent studies show that the traceability of energy regions in the spectro-temporal representation may well link to features relevant for recognition ([12] and [13]), although they also indicate the relative importance of such regions to recognition may be more complex. Lewicki et. al. ([14]) represent yet another reason for looking at biological systems. They compared filters as derived from experimental data (known as revcor filter) to mathematical filters (wavelet and Fourier) and showed up to a three-fold increase of coding efficiency of speech signals by revcor filters over Fourier or wavelet based systems for noisy signals. The subject of coding efficiency is however beyond the scope of this paper.

5. REFERENCES

[1] George Zweig, "Finding the impedance of the organ of Corti," *The Journal of the Acoustical Society of America*, vol. 89, no. 3, pp. 1229–1254, 1991.

[2] Trevor J Cooke and Odette Scharenborg, "The Interspeech 2008 Consonant Challenge," in *Proceedings of Interspeech 2008*, Jan. 2008.

[3] H Hirsch and D Pearce, "The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," in *Automatic Speech Recognition: Challenges for the Next Millenium*, 2000.

[4] R Hut, M M Boone, and A Giesolf, "Cochlear modeling as time-frequency analysis tool," *Acta Acustica united with Acustica*, vol. 92, no. 4, pp. 629–636, 2006.

[5] Dennis Gabor, "Theory of communication," *Journal IEE*, vol. 93, no. 26, pp. 429–457, 1946.

[6] Kamil Wojcicki, "Htk mfcc matlab," 11 2011.

[7] Stéphane G Mallat, *A wavelet tour of signal processing*, Academic Press, Burlington, MA, United States of America, 2009.

[8] V Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," *Acta Acustica United with Acustica*, vol. 88, no. 3, pp. 433–442, 2002.

[9] T Herzke and V Hohmann, "Improved numerical methods for gammatone filterbank analysis and synthesis," *Acta Acustica United with Acustica*, vol. 93, pp. 498–500, 2007.

[10] H Duifhuis, HW Hoogstraten, van Netten, Rob J Diependaal, and W Bialek, "Modelling the cochlear partition with coupled Van Der Pol oscilators," in *Peripheral Auditory Mechanisms*, Jont B Allen, S T Neely, and A Tubis, Eds., pp. 290–297. Springer Verlag, Berlin, 1986.

[11] George Zweig and Christopher A Shera, "The origin of periodicity in the spectrum of evoked otoacoustic emissions," *The Journal of the Acoustical Society of America*, vol. 98, no. 4, pp. 2018–2047, Jan. 1995.

[12] F Li, A Menon, and J B Allen, "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2599–2610, 2010.

[13] A Kapoor and J B Allen, "Perceptual effects of plosive feature modification," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 478–491, 2012.

[14] Evan C Smith and Michael S Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, Feb. 2006.