2005 International Symposium on Nonlinear
Theory and its Applications (NOLTA2005)
Bruges, Belgium, October 18-21, 2005

Sparse Time-Frequency Analysis of Speech Signals

Albert Kern†, Oliver Nagy‡, and Ruedi Stoop†

†Institute for Neuroinformatics, University / ETH Zürich, 8057 Zürich
Email: albert@ini.phys.ethz.ch, ruedi@ini.phys.ethz.ch
‡Department of Electrical Engineering, ETH Zürich
Email: nagyo@ee.ethz.ch

Abstract—To obtain a sparse time-frequency representation of a speech signal, which still reflects the evolution of its characteristics in time, consists a great problem. Despite its disadvantages, the spectrogram is traditionally used for this purpose. In this contribution we examine the potential of 1) atomic decomposition methods and 2) time-frequency distributions with respect to speech signals.

1. Introduction

Speech signals are highly transient, consisting of voiced segments exhibiting a well-defined frequency structure, which are separated by noise bursts (consonants). It is therefore a great problem to devise a representation for a speech signal which 1) reflects the evolution of its frequency contents, 2) is sparse and robust with respect to noise, and 3) conveys all relevant time-frequency information. If such a time-frequency representation of a speech signal has been found, auditory scene analysis can be performed by using an appropriate clustering algorithm.

Traditionally, the spectrogram has been used for obtaining a time-frequency representation of speech signals. However, the disadvantage of the spectrogram consists in the inherent trade-off between time and freqency resolution, which is determined by the window function used in the short-time Fourier transform. If a sufficient time resolution is to be maintained, this has to be paid for by a blurring of the frequency components. Although it is impossible, due to Heisenberg's uncertainty principle, to simultaneously locate a signal component in time and frequency (see e.g. [1]), alternative time-frequency representation methods, that come closer to this Heisenberg limit, have been proposed [2, 3, 4]. However, each of these methods also has its advantages and drawbacks, depending on the class of signals to which they are applied.

In this contribution, different types of time-frequency representation methods for analyzing speech signals are examined. In general, two classes of time-frequency representation methods exist: 1) Atomic decomposition methods and 2) time-frequency distributions. In the first case, the signal is decomposed into a set of basis functions, which are chosen from an overcomplete dictionary. Preferably, the basis functions are concentrated in time and frequency. Optimal results are obtained if, in addition, these basis functions capture essential characteristics of the signal. Therefore, the selection of an optimal dictionary requires a-priori knowledge about the time-frequency structure of the signal. Wavelet decomposition, matching and basis pursuit methods belong to this class.

Alternatively, the distribution of the signal energy in time and frequency can be considered. This leads to the class of time-frequency distributions, where the bilinear distributions play a major role. The Wigner-Ville distribution, which originally has been proposed in the context of theoretical problems in quantum mechanics [5] is the earliest, and most well-known, member of this class.

2. Atomic Decomposition Methods

Suppose we are given a signal $f(t) \in L^2(\mathbb{R}$ (e.g. recorded speech), which is decomposed into a set of functions $\phi_i(t)$, which are well localized in time and frequency,

$$f = \sum_{\phi_i \in \Gamma} \alpha_i \phi_i, \tag{1}$$

The "time-frequency atoms" ϕ_i, are chosen from a dictionary Γ. Γ may constitute a basis of $L^2(\mathbb{R})$, or it may be overcomplete. In the latter case, several subsets of Γ constitute a basis for $L^2(\mathbb{R})$, so that different selections of $\phi_i \in \Gamma$ are possible. When using an appropriate basis selection method, this freedom may allow for a better, i.e. sparser, representation of the signal $f(t)$.

A *sparse* representation of $f(t)$ is obtained if only a small number m of coefficients α_i assume an appreciable value,

$$f = \sum_{i=1}^{n} \alpha_i \phi_i + R^{(m)}(f), \quad \phi_i \in \Gamma, \tag{2}$$

where $R^{(m)}(f)$ denotes the residual. This decomposition should reflect the time-frequency content of the signal f.

A graphical representation is obtained by weighting the time-frequency support of ϕ_i by the coefficient $|\alpha_i|$; this provides a measure for the "energy of interaction" between the signal f and the time-frequency atom ϕ_i. An optimal time-frequency representation of the signal would thus allow to "read off" its fundamental constituents – provided that these are reflected by some time-frequency atoms $\phi_i \in \Gamma$.

One way to fast and efficiently obtain a signal decomposition (2) is to use wavelet analysis, where the dictionary Γ consists of wavelet functions, that are derived from

a mother wavelet by dilations and translations. However, since speech signals do not exhibit scaling properties, the application of the wavelet decomposition to speech signals proves less valuable: For example, the formant structure of voiced speech segments demands a fine frequency resolution, even at high frequencies, while the resolution of the wavelet decomposition rapidly decays at higher frequencies [6]. Therefore, other dictionaries Γ, containing atoms ϕ_i that are better adapted to the properties of speech signals, must be used. In the following section, two methods for signal decompositions using arbitrary dictionaries are introduced.

In 1993, Mallat and Zhang [7] proposed an algorithm for stepwise obtaining a sparse approximation (2). We start with $f^0 = 0$ for the initial approximation of the signal f, and $R^{(0)} = f$ for the residual. At each step k. the basis function $\phi_k \in \Gamma$ is chosen such that the modulus of $\alpha_k =< R^{(k-1)}, \phi_k >$ is maximal (where $R^{(k-1)} = f - f^{k-1}$ denotes the residual at step $k-1$ and $< \cdot, \cdot >$ is the scalar product). The procedure may either be stopped after n steps (where n is determined in advance), or if $\|R^n\|$ falls below a predetermined threshold.

For non-orthogonal dictionaries Γ, however, unsatisfactory results may occur. Since the algorithm is myopic, it may choose wrongly in the first few iterations. As a consequence, the resulting error must be corrected in many subsequent iterations, leading to a non-optimal signal decomposition. For resolving this problem, sophisticated methods like *back-fitting* [7] and orthogonal matching pursuit [8, 9] have been devised.

The principle of basis pursuit [10, 11] is to find a representation of the signal f such that the decomposition coefficients α_k have minimal ℓ^1-norm,

$$\min \|\alpha\|_1 = \sum_k |\alpha_k| \quad \text{for} \quad \sum_k \alpha_k \phi_k = f. \tag{3}$$

In contrast to matching pursuit, basis pursuit is a global optimization principle, which leads to a convex optimization problem. While matching pursuit builds up the decomposition step by step, basis pursuit starts with the full model and improves it by swapping atoms for more useful ones. Therefore, in comparison to basis pursuit, matching pursuit is sub-optimal. The minimization of the ℓ^1-norm constitutes a computationally demanding problem, which can be efficiently implemented by using recent advances in linear programming theory (interior point methods, [12]).

3. Time-Frequency Energy Distributions

3.1. Bilinear Distributions

The Wigner-Ville distribution (WVD)

$$W(t,\omega) = \frac{1}{2\pi}\int f^*\left(t-\frac{\tau}{2}\right) f\left(t+\frac{\tau}{2}\right) e^{-i\tau\omega} d\tau, \tag{4}$$

$$= \frac{1}{2\pi}\int \hat{f}^*\left(\omega+\frac{\theta}{2}\right) \hat{f}\left(\omega-\frac{\theta}{2}\right) e^{-it\theta} d\theta, \tag{5}$$

plays a central role among the time-frequency energy distributions. It leads to optimal time-frequency resolution for linear chirp signals. Since the WVD is bilinear, however, if different signal components are present, or if the frequency modulations are nonlinear, cross terms are generated. For broad-band, multi-component signals, like speech, this makes the obtained time-frequency representation difficult to interpret.

Moreover, the WVD may attain negative values (again with the exception of Gaussian linear chirps), which renders its interpretation as an energy density difficult. It has been shown [3] that the occurrence of negative values is closely linked to the emergence of the (rapidly oscillating) cross terms. The construction of purely positive distributions (see below) may thus serve to reduce cross terms.

For finite duration signals, the WVD leads to boundary effects. A windowed version of the WVD,

$$W_{ps}(t,\omega) = \frac{1}{2\pi}\int h^*\left(-\frac{\tau}{2}\right) f^*\left(t-\frac{\tau}{2}\right) h\left(\frac{\tau}{2}\right) f\left(t+\frac{\tau}{2}\right) e^{-i\tau\omega} d\tau,$$

$$= \int S_f^*(t,\omega-\theta) S_f(t,\omega+\theta) d\theta, \tag{6}$$

called the *Pseudo-Wigner distribution* (PWD), resolves this problem. $h(t)$ denotes the window function (e.g. Hanning window), and $S_f^*(t,\omega)$ is the short-time Fourier transform (STFT) of $f(t)$.

In order to reduce the emergence of cross-terms, several methods have been devised. The *kernel method* consists in convolving the WVD by a suitably chosen kernel, $\phi(\theta,\tau)$, resulting in a new bilinear distribution,

$$C(t,\omega) = \frac{1}{4\pi^2}\iiint f^*(u-\frac{\tau}{2}) s(u+\frac{\tau}{2}) \phi(\theta,\tau) e^{-i\theta\tau - i\tau\omega + i\theta u} du d\tau d\theta. \tag{7}$$

It has been shown [2] that every bilinear time-frequency distribution can be expressed in this way (where the kernel $\phi(\theta,\tau)$ has to satisfy several conditions). In particular, the kernel may be chosen in such a way as to reduce the emergence of cross terms. A frequently used example is the Choi-Williams kernel [13]. However, with respect to speech signals the Choi-Williams distribution proves still unsatisfactory. The S-method and the class of positive distributions appear to have a larger potential for the analysis of speech signals.

3.2. S-Method

The S-Method [4] is obtained by convolving the Pseudo-Wigner distribution (6) by the window function $P(\theta)$,

$$SM(t,\omega) = \int P(\theta) S_f^*(t,\omega-\theta) S_f(t,\omega+\theta) d\theta. \tag{8}$$

Evidently, the choice $P(\theta) = \delta(\theta)$ reduces $SM(t,\omega)$ to the STFT; on the other hand, using $P(\theta) \equiv 1$ leads to the Pseudo-Wigner distribution (6). Optimal results may be obtained by choosing

$$P(\theta) = \begin{cases} 1 & \forall \quad |\theta| \le L_p \\ 0 & \text{else,} \end{cases} \tag{9}$$

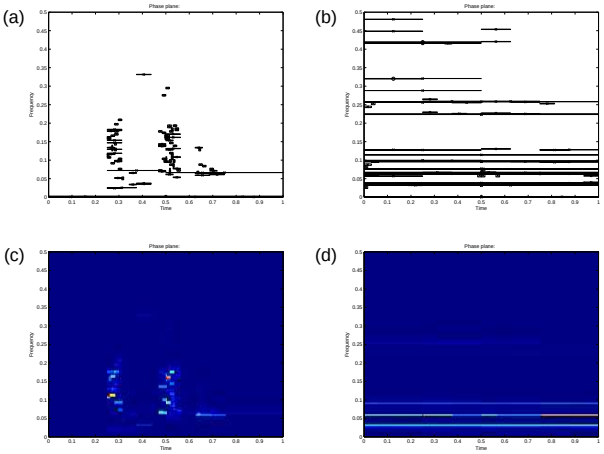

Figure 1: Matching pursuit method (local cosine-packet dictionary, 100 basis functions) applied to the Japanese word “arigato” (panels (a) and (c)) and to a superposition of the vowels /u/ and /e/ (panels (b) and (d)). (a) and (b): Time-frequency supports; (c) and (d): support weighted by decomposition coefficients.

where L_p denotes the length of the rectangular window $P(\theta)$, which may be adaptively determined, depending on the signal properties.

3.3. Positive Time-Frequency Distributions

It has already been stated that the emergence of cross-terms is linked with the non-positivity of a time-frequency distribution. The construction of positive time-frequency distributions (PTFD) may therefore lead to sparser representations.

It has been pointed out [4] that the searched optimal PTFD of a not too violently varying signal $f(t)$ can be represented by a convolution of its spectrogram $F_f(t,\omega)$ with WVD of the analysis window, $W(t,\omega)$,

$$F_f(t,\omega) \approx \iint P(\tau,\theta)W(t-\tau,\omega-\theta)d\tau d\theta. \qquad (10)$$

As it is necessary that $W(t,\omega)$ be non-negative everywhere, a Gaussian function must be chosen for the analysis window. The PTFD $P(t,\omega)$ is then recovered by applying a deconovolution algorithm [14].

4. Application to Speech Signals

Different time-frequency representation methods have been applied to short speech segments, consisting of a) the superposition of two vowels and b) a short Japanese word (“arigato”).

Fig. 1 displays the results obtained from the matching pursuit method. Fig. 1 has to be compared with the results obtained from time-frequency energy distributions (Figs. 2 and 3). As a first observation, we note that the

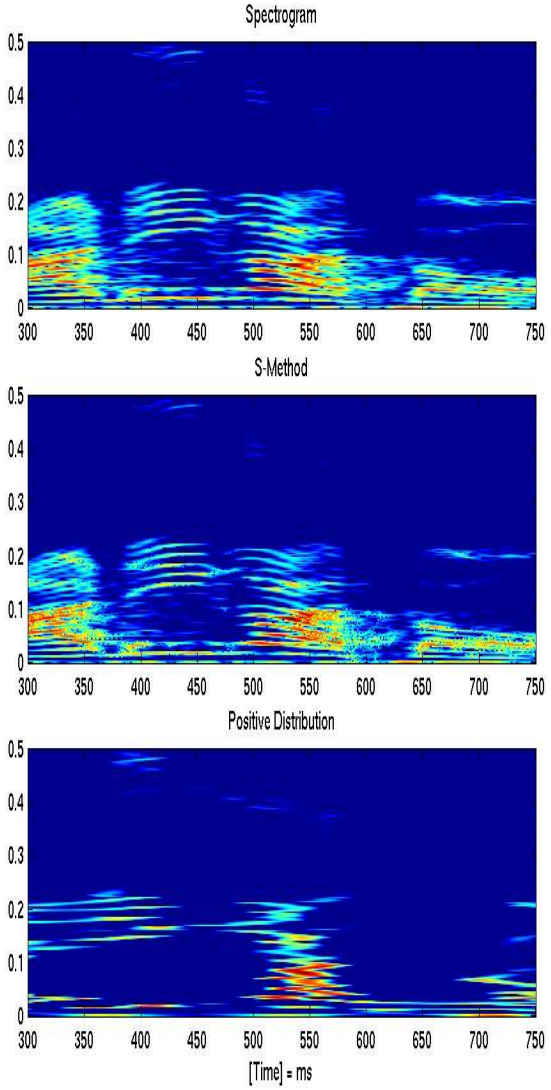

Figure 2: Time-frequency representations of the Japanese word “arigato”.

time-frequency distributions reveal a richer signal structure than the matching pursuit method (Fig. 1 (a), (c) vs. Fig. 2). This can be explained by the choice of the dictionary: Since we chose a local cosine packet dictionary (which leads to fast convergence of the matching-pursuit algorithm), frequency modulations of harmonic speech components are not matched by members of the dictionary. Instead, a superposition of a large number of low-energy time-frequency atoms is needed in order to take account of such components. Since the number of atoms used in the representation was restricted, however, these low-energy components were discarded completely. If this maximal number was increased, in the resulting time-frequency representation the characteristics of the signal (frequency modulation of single harmonic components) would still remain masked by the emerging large number of constant-frequency atoms.

This problem could be resolved by using a dictionary

consisting of frequency-modulated atoms (chirplets) [15]. As the number of elements of such a dictionary is vastly increased in comparison to the previous situation, matching pursuit encounters a considerable computational demand.

When comparing the different time-frequency distribtion methods (spectrogram, S-method and PTFD) we observe that the PTFD leads to the highest time-frequency resolution. In particular, the beats originating from constructive and destructive interference when superimposing vowels /u/ and /e/ (and which are perceived when listening to the sample), are clearly revealed by the PTFD.

In order to improve the performance of the matching pursuit method, we propose that the PTFD or the S-method is computed first. This provides information about location and modulation properties of different harmonic speech components. This information may then be used to efficiently search for time-frequency atoms in a large dictionary, which contains elements that account for frequency and amplitude modulation.

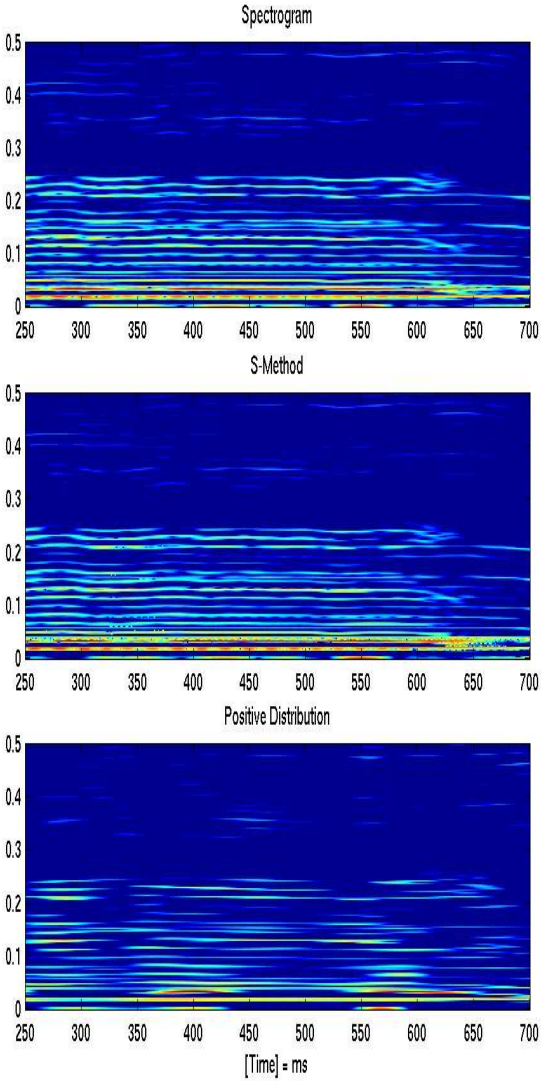

Figure 3: Time-frequency representations of a superposition of the two vowels /u/ and /e/.

References

[1] K. Gr¨ochenig, *Foundations of Time-Frequency Analysis* (Birkh¨auser, 2001).

[2] L. Cohen, *Time-Frequency Analysis* (Prentice Hall, Upper Saddle River, 1995).

[3] P. Flandrin, *Time-Frequency/Time-Scale Analysis*, (Academic Press, 1999).

[4] B. Boashash (Ed.), *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*, (Elsevier, 2003).

[5] E.P. Wigner, "On the quantum correction for thermodynamic equilibrium", *Physical Review*, vol. 40, pp. 749–759, 1932.

[6] Note that the wavelet decomposition reflects the behaviour of the signal $f(t)$ at different *scale*, rather than frequency. However, as a simple approximation, the frequency may be taken as the inverse of scale.

[7] S.G. Mallat, Z. Zhang, "Matching pursuits with time-frequency dictionaries", *IEEE Trans. Signal Proc.*, vol. 41, pp. 3397–3415, 1993.

[8] G. Davis, S. Mallat, Z. Zhang, "Adaptive time-frequency decompositions", *Optical Engineering*, vol. 33, pp. 2183–2191, 1994.

[9] L. Rebollo-Neira, D. Lowe, "Optimized orthogonal matching pursuit approach", *IEEE Signal Proc. Letters*, vol. 9, pp. 137–140, 2002.

[10] S.S. Chen, D.L. Donoho, M.A. Saunders, "Atomic decomposition by basis pursuit", *SIAM Rev.*, vol. 43, pp. 129–159, 2001.

[11] D.L. Donoho, M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization", *Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, pp. 2197–2202, 2003.

[12] I.J. Lustig, R.E. Marsten, and D.F. Shanno, "Interior point methods for linear programming: Computational state of the art", *ORSA Journal on Computing*, vol. 6, pp. 1–14, 1994.

[13] H.I. Choi, W.J. Williams, "Improved time-frequency representations of multicomponent signals using exponential kernels", *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. 37, pp. 862–871, 1989.

[14] D. Snyder, T. Schultz, J. O'Sullivan, "Deblurring subject to nonnegativity constraints", *IEEE Trans. Signal Proc.*, vol. 40, pp. 1143–1150, 1992.

[15] R. Gribonval, "Fast matching pursuit with a multiscale dictionary of Gaussian chirps", *IEEE Trans. Signal Proc.*, vol. 49, pp. 994–1001, 2001.