MATCHING PURSUIT DECOMPOSITIONS OF NON-NOISY SPEECH SIGNALS USING SEVERAL DICTIONARIES

Bob L. Sturm and Jerry D. Gibson

Department of Electrical and Computer Engineering
University of California, Santa Barbara, Santa Barbara, CA 93106 USA
http://www.mat.ucsb.edu/ b.sturm

ABSTRACT

Matching pursuit (MP) provides a way to expand signals in terms of any set of time-limited functions, or atoms, called a dictionary. These decompositions are finding use in signal analysis and coding. It has been shown that a dictionary should be designed carefully, but its effects on decomposition have not been studied in detail. We look at the effects of dictionaries on the decomposition of non-noisy speech signals using MP, by five dictionaries. It is found that Gabor atoms work sufficiently well, and have fewer adverse effects in reconstruction compared to the other dictionaries. For a reconstruction to sound perceptually close to the original, a rate of 3000 atoms per second (aps) on average is required. At rates as low as 400 aps the speech remains intelligible. Finally, the use of decompositions to visualize time-frequency distributions of speech is explored.

1. INTRODUCTION

Relatively recent work has explored expanding digital signals in terms of functions that are more representative of the signal than other basis functions. One such method, matching pursuit (MP), iteratively finds best matches of a signal to vectors in a usually highly redundant and over-complete set of time-limited functions, or atoms, called a dictionary [1]. The selection is based on maximizing the inner product of the signal and vectors in the dictionary, which minimizes the squared error of the reconstruction. The signal can then be represented as a linear combination of N scaled atoms and a residual:

$$f(t) = \sum_{n=0}^{N-1} \langle R^n f, g_{\gamma_n} \rangle g_{\gamma_n} + R^N f \tag{1}$$

where $g_{\gamma_n}(t)$ is a unit-norm vector in the dictionary, which is indexed by γ_n. After each match, the parameters of the selected atom are stored and the process repeats on the residual signal until its energy passes a specified minimum or a required number of atoms have been found. Each atom is characterized by several parameters, such as size, center time, frequency, phase, and norm. The resulting representations can be more sparse and flexible than other expansions, but at the price of efficiency, convergence, and increased computation.

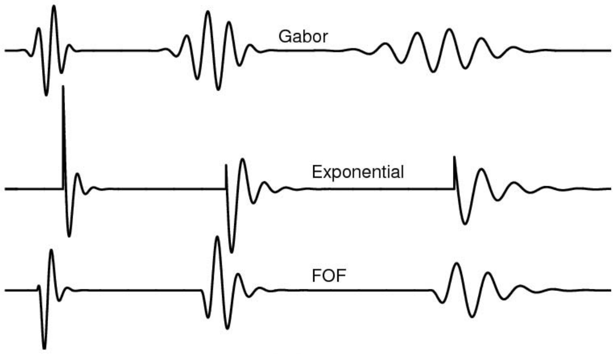

Fig. 1. Examples of the dictionary atoms used in this study.

MP decompositions (MPD) have been used to encode and reconstruct a signal [2], aid in feature extraction, signal analysis, and classification [3, 4], and transformation and visualization [5]. By superimposing the Wigner-Ville distribution (WVD) of each atom selected by a decomposition, a time-frequency distribution (TFD) can be created that has smaller time and frequency support than that given by other methods, such as the short-term Fourier transform (STFT).

Mallat and Zhang [1], and Davis [6], have studied the theoretical behaviors of the decomposition process, and the importance of the dictionary. Others have researched optimal dictionary sizes for decompositions [7]. Some have even suggested that particular atoms are more efficient for some signal types, such as damped sinusoids for speech [8]. Despite this growing research, there exists little exploration in the actual effects of different dictionaries on signal decomposition. Given monophonic, non-noisy speech, how efficiently can it be represented and how accurately can it be reconstructed using a given dictionary? In this paper we present the results of decomposing several speech signals with five different dictionaries using non-orthogonal MP.

Research supported in part by NSF grants No. DGE-0221713 (IGERT in Interactive Digital Multimedia) and No. CCF-0429884.

1-4244-0469-X/06/$20.00 ©2006 IEEE

2. IMPLEMENTATION AND SPEECH EXAMPLES

To investigate the effects of dictionaries on MPD of speech, we have used the LastWave (LW) software package [9], with the MP implementation by Gribonval, Bacry, and Abadia. There are numerous atom types available in this software, including Gabor atoms (modulated Gaussians) [10], complex exponentials, and "Fonction d'onde Formantique" (FOF) [11]. Figure 1 shows examples of each type of atom.

Inspecting the code of LW, one can find that for an input signal of N samples, a dictionary is created that virtually has $26N$ vectors for each atom type. Atom sizes are limited to thirteen powers of 2, from 4 to 16384 samples, which is imposed to speed the algorithm. When searching for best matches, atom times are skipped by one quarter the atom size. The number of frequencies each atom can have is one half their size in samples. The criteria for atom selection is determined by the largest inner product of the signal with the dictionary.

Four short speech signals from different speakers were selected for this study, taken from "Books on CD." Each signal was reduced to one channel, lowpass filtered, and downsampled to 8 kHz, with 16-bit quantization. Decompositions of each signal were found using three homogeneous, and two hybrid dictionaries: Gabor atoms, complex exponentials, FOF, and unions of the Gabor atoms with the other two. These atom types were selected because their inner products can be computed analytically [9]. Half the speech signals were from male speakers ("speech 2,", "3"). One of the female speakers produced a pathological voice ("speech 4").

In each MPD, the entire signal was decomposed at once, as opposed to using a windowed approach. The reason this was done was first to see how the algorithm distributes the atoms, second to eliminate redundancy created by overlapping windows, and third to reduce the possibility of the algorithm choosing atoms that do not exist within the signal, e.g. block effects.

3. COMPARISONS OF DICTIONARIES

Figure 2 shows the decay of residual energy for each speech signal using a Gabor dictionary as a function of the average atom rate. The average number of atoms per second (aps) is employed because when decomposing an entire signal at once the actual rate of atoms varies over time depending on the concentration of energy in the signal.

The rate of residual energy decay approaches an asymptotic limit that depends only on the dictionary, and not the signal [6]. After an initial quick descent, when the dictionary is very similar to the signal, the residual decay rate slows to that limit–approximately -8 dB per 1000 aps for the Gabor dictionary. The slow decay rate of "speech 3" can be explained by the presence of equalization; its spectrum was flatter than the others, thus requiring more atoms.

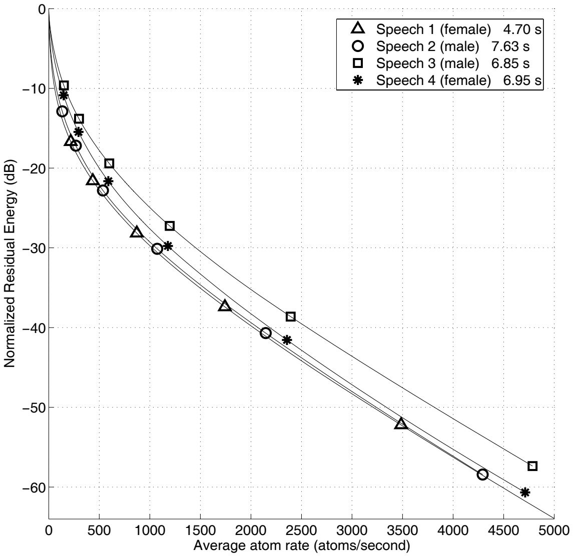

Fig. 2. Residual energy of four speech signals as a function of average atom rate for Gabor dictionary. Duration of each signal is given in legend.

Figure 3 plots the decay of residual energy for two speech signals for each dictionary, as a function of atom rate. The same behavior is observed for the other speech signals. It is clear here that the dictionary of complex exponentials is the least efficient at representing the signals. For "speech 1," at a residual level of -50 dB, over 1000 more aps are needed for the exponential than for the Gabor dictionary. This is not surprising since exponential atoms are asymmetric and discontinuous, requiring more atoms to correct for these characteristics. The asymptotic decay rate of the complex exponential dictionary is about -7 dB per 1000 aps.

Both hybrid dictionaries performed better than the three homogeneous ones, which is predictable since increasing the size of a dictionary usually gives a more sparse decomposition [7]. The dictionary that combines Gabor and FOF atoms performed the best, though minimally so compared with the homogeneous Gabor dictionary. At a residual level of -50 dB, this difference is only about 250 aps.

3.1. Reconstructions

By listening to the reconstructions, it was found that a minimum rate of 3000 aps for each dictionary except the exponentials, is required for near transparency. The exponential dictionary required over 1000 aps more to reach a similar level of quality. From figure 3 it can be seen that the residual energy is at -50 dB at these rates. When considering only the intelligibility of the speech, it was found that the average atoms rates could be as low as 400 aps.

By decomposing a signal in its entirety, as opposed to decomposing time-limited portions of it, a constant rate of atoms cannot be guaranteed. When using the inner product as the choice function, MP will naturally choose more atoms in

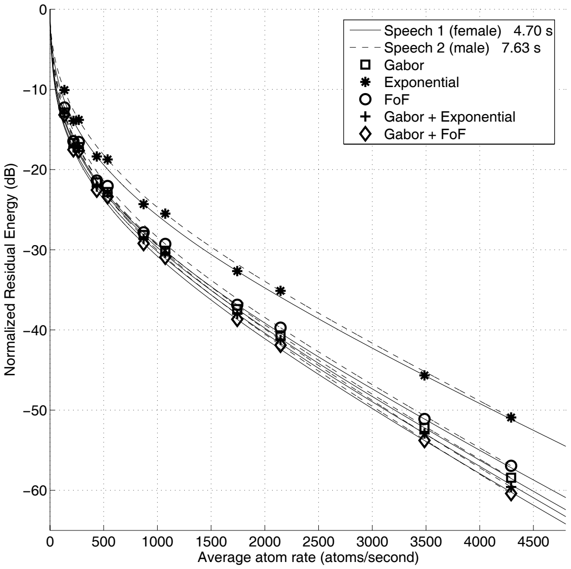

Fig. 3. Residual energy of two speech signals as a function of average atom rate for the five dictionaries.

regions of high energy, to the detriment of other low energy, but possibly perceptually significant, regions. The effects of this can be heard; at times the reconstruction sounds incomplete. These effects can be resolved by using a windowed approach to the decomposition, but at a price of redundancy in the analysis.

3.2. Time-Frequency Distributions

Figure 4 compares two TFDs of "speech 1," one acquired from the superposition of the WVD of each atom selected in the MPD using Gabor atoms (middle), and the other created using a narrowband STFT (bottom). The speech signal is seen at top with a voiced/unvoiced profile. One can clearly see that though the WVD of an atom gives a superior TFD of that atom, in the sense that its time and frequency support are minimized, doing so for a collection of atoms found from a decomposition of a natural signal may not provide an accurate representation of the signal energy distribution.

The arrows show instances of where the decomposition displays curious behaviors: (1) points to a time that shows energy in the WVD, but in the time domain has minimal energy. Though the combination of the atoms in this case sums to zero, this is not reflected by the WVD, as it is in the STFT. This "failure" has been noted before [12]. (2) points to the sibilance of the word "she" decomposed as several short high frequency atoms. The MPD has found a mass of atoms that attempt to precisely characterize that realization of noise. (3) points to a fine speech structure missing entirely from the WVD. (4) points to a dipthong approximated by several short duration atoms. All these effects are seen in WVDs using the other dictionaries, but at times less pronounced. For instance the WVD using FOFs exhibits sharper onsets than can be provided by symmetric Gabor atoms.

It may be surprising that even with such large differences between the two TFDs, the reconstructions sound the same. Though the WVD doesn't match the STFT, when the signal is reconstructed the atoms interfere in just the right ways to create silences, such as that pointed to by (1).

4. CONCLUSIONS

We have experimentally compared the performance of five different dictionaries for non-orthogonal MPDs of four non-noisy speech signals. It has been shown that the performance of dictionaries containing Gabor atoms work sufficiently well at representing the signal energy. These decompositions require on average 1000 aps less than decompositions using exponential atoms at a residual energy of about -50 dB. Furthermore, diversifying the dictionary with FOF atoms reduces the aps required for the same residual energy by only 250. Each dictionary except for the exponential gave a residual energy decay rate of around -8 dB per 1000 aps.

Summing the WVD of individual atoms to visualize the TFD of a signal might be misleading. Not only can fine frequency structures disappear, but energy is shown where none exists in the signal. These problems however can be addressed by using more diverse dictionaries (e.g. chirps), using modified MP algorithms, such as the high resolution MP [3], or quite simply modifying the image using the energy envelope of the signal.

Though MP aims at finding sparser representations of signals, it does so at the price of convergence, efficiency, and accuracy in the frequency domain, for each of the dictionaries tested. For nearly transparent reconstructions of these speech signals at 8 kHz, requiring 3000 aps, this implies a rather low average sparsity of about 0.375 atoms per sample. Structures of speech, such as band-limited shaped noise, become masses of atoms each having center times, durations, amplitudes, frequencies, and phases–in short an unnecessary explosion of data.

Further directions of research will include comparing different decomposition strategies, e.g. windowed analysis, atoms plus noise; applying psychoacoustic principles to streamline the algorithm and results; determining the sensitivity of each atom parameter on the reconstruction; and specializing the dictionary for representing speech.

5. REFERENCES

[1] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Sig. Proc.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[2] P. Vera-Candeas, N. Ruiz-Reyes, M. Rosa-Zurera, D. Martinez-Munoz, and J. Curpián-Alonso, "New matching pursuit based sinusoidal modelling method for audio coding," *IEEE Proc.-Vis. Image Signal Process.*, vol. 151, no. 1, pp. 21–28, 2004.

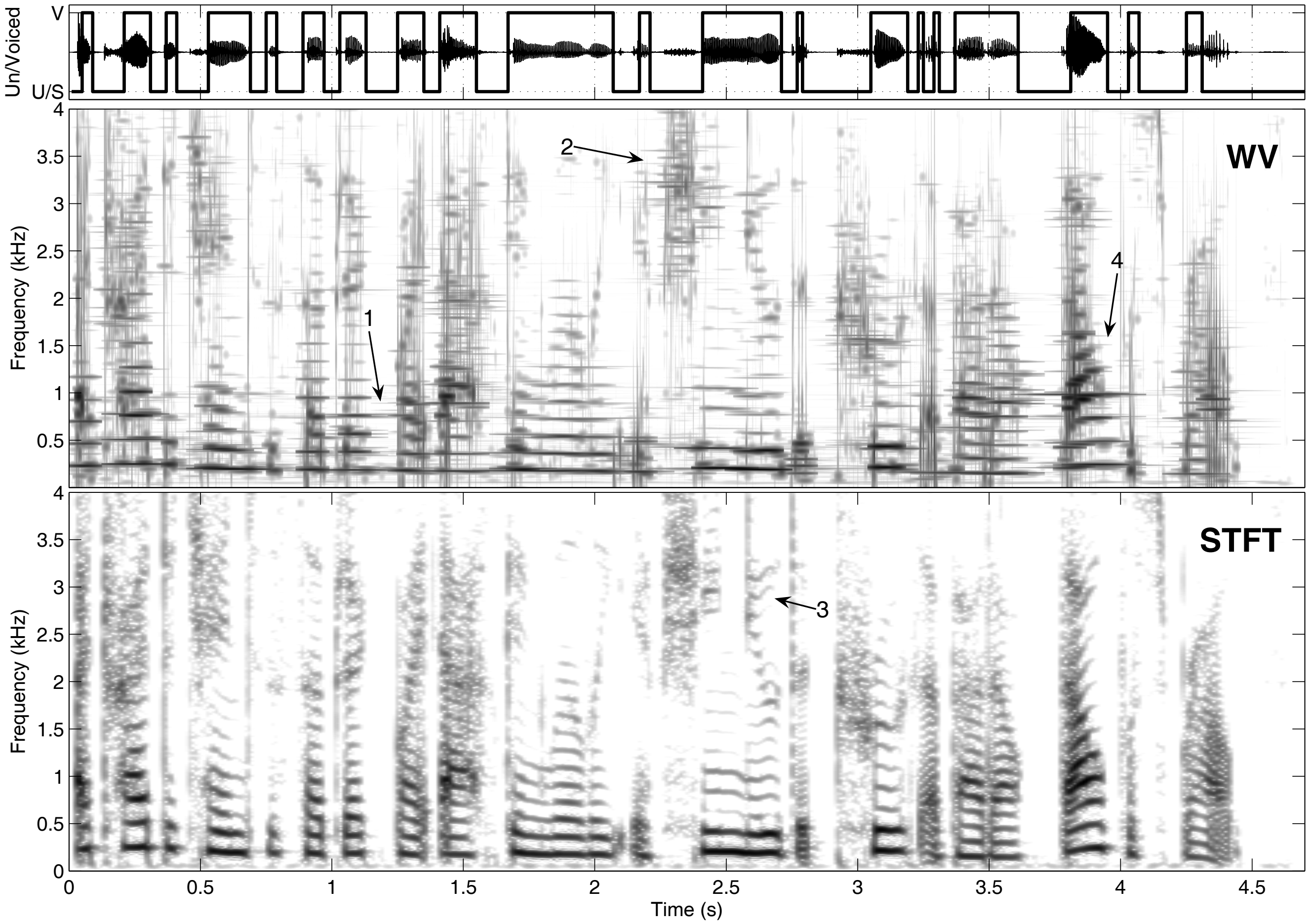

Fig. 4. The WVD of the decomposition using Gabor atoms (middle), and narrowband STFT (bottom), are shown for "speech 1" with voiced/unvoiced profile (top). Sentence is: "I tried to change the subject by asking Lilly if she knew the truth about alfalfa sprouts." Arrows are discussed in text.

[3] S. Jaggi, W. Carl, S. Mallat, and A. Willsky, "High resolution pursuit for feature extraction," *Technical report, MIT, November*, 1995.

[4] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson, "Discrimination of pathological voices using a time-frequency approach," *IEEE Trans. Biomedical Eng.*, vol. 52, no. 3, pp. 421–430, 2005.

[5] G. Kling and C. Roads, "Audio analysis, visualization, and transformation with matching pursuits," in *Proc. of the COST G-6 Conference on Digital Audio Effects (DAFX-04)*, 2004.

[6] G. Davis, *Adaptive Nonlinear Approximations*, Ph.D. thesis, New York University, 1994.

[7] Q. Liu, Q. Wang, and L. Wu, "Size of the dictionary in matching pursuit algorithm," *IEEE Trans. Sig. Proc.*, vol. 52, no. 12, pp. 3403–3408, 2004.

[8] M. Goodwin and M. Vetterli, "Matching pursuit and atomic signal models based on recursive filter banks," *IEEE Trans. Sig. Proc.*, vol. 47, no. 72, pp. 1890–1902, 1999.

[9] E. Bacry, "Lastwave software [online]," www.cmap.polytechnique.fr/ bacry/LastWave/.

[10] D. Gabor, "Acoustical quanta and the theory of hearing," *Nature*, vol. 159, no. 4044, pp. 591–594, 1947.

[11] X. Rodet, "Time-domain formant-wave function synthesis," in *Spoken Language Generation and Understanding*, J. C. Simon, Ed., pp. 429–441. D. Reidel, New York, 1980.

[12] R. Gribonval, P. Depalle, X. Rodet, E. Bacry, and S. Mallat, "Sound signals decomposition using a high resolution matching pursuit," in *Proc. Int. Computer Music Conference*, 1996.