International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:5, 2015

A Simple Adaptive Atomic Decomposition Voice Activity Detector Implemented by Matching Pursuit

Thomas Bryan, Veton Kepuska, Ivica Kostanic

***Abstract*—**A simple adaptive voice activity detector (VAD) is implemented using Gabor and gammatone atomic decomposition of speech for high Gaussian noise environments. Matching pursuit is used for atomic decomposition, and is shown to achieve optimal speech detection capability at high data compression rates for low signal to noise ratios. The most active dictionary elements found by matching pursuit are used for the signal reconstruction so that the algorithm adapts to the individual speakers dominant time-frequency characteristics. Speech has a high peak to average ratio enabling matching pursuit greedy heuristic of highest inner products to isolate high energy speech components in high noise environments. Gabor and gammatone atoms are both investigated with identical logarithmically spaced center frequencies, and similar bandwidths. The algorithm performs equally well for both Gabor and gammatone atoms with no significant statistical differences. The algorithm achieves 70% accuracy at a 0 dB SNR, 90% accuracy at a 5 dB SNR and 98% accuracy at a 20dB SNR using 30d B SNR as a reference for voice activity.

***Keywords*—**Atomic Decomposition, Gabor, Gammatone, Matching Pursuit, Voice Activity Detection.

I. Introduction

VOICE activity detection is a vital process for automatic speech recognition systems like SIRI for the Iphone or “Ok Google” for Android phones. These applications work poorly or not at all in high noise environments. Activating a turn signal in a car, for example, may generate at “tick-tock” sound that can cause the system to fail. The first line of defense for automatic speech recognition systems is the VAD. Virtually all speech processors contain a VAD algorithm that attempts to estimate portions of signals that correspond to speech while ignoring sources of ambient noise. Automatic speech recognition systems are very sensitive to noise, so a robust VAD is a critical component in optimizing the systems accuracy. If the VAD falsely detects noise as speech, the speech recognition algorithm is given bad input. Conversely, if the VAD does not detect a valid speech signal, then the real data is missed which also results in poor performance.

The background noise is not typically Additive White Gaussian Noise (AWGN) for real world speech recognition applications. However, a first step in designing a VAD is to measure its performance with AWGN to provide a baseline of performance. Some types of noise are related to AWGN such as pink noise, and from the central limit theorem when there are multiple sources of background noise, the distribution tends to be Gaussian, although not necessarily white. Enhancements to the baseline VAD may subsequently be introduced to combat other types of noises.

Gabor proposed that speech could be represented as “quanta of information” in the time-frequency plane [1]. Gabor recognized the time-frequency uncertainty problem of Fourier methods was similar to the Heisenberg uncertainty of Quantum physics. He proposed an atom composed of a cosine wave modulated by a Gaussian pulse. This fundamental “atom” can be elongated on contracted in time and amplitude to make a dictionary of atoms. Gabor believed that speech could be modeled as a superposition of these atoms in the time domain. In doing so, he theorized the fundamental information in the speech signal is captured in the temporal and amplitude coefficients of the atoms.

Mallat and Zhang introduced the matching pursuit algorithm which finds the best inner products between data and dictionary atoms for all possible shifts of each atom [7]. The matching pursuit algorithm isolates structures in the data that are coherent with atoms from a given dictionary which is typically overcomplete. Linear expansions such as Fourier basis and wavelets are not flexible enough to represent complex data like speech that exhibit significant variation in the time-frequency plane [7]. For example, short impulses might best represent plosives in speech that are concentrated in time, whereas long duration formant frequencies may be better represented by long duration sinusoids with narrow frequency support. Fourier basis signal are not able to localize time, whereas multiresolution based discrete wavelet transforms are not able to localize high frequency components. Wigner or Cohen’s class distributions are able to simultaneously localize time and frequency, but contain interference terms. Atomic decomposition by matching pursuit is able to provide a clear picture of complex data in the time-frequency plane.

Matching pursuit is an attractive approach for isolating signals that have a high Peak to Average Power Ratio (PAPR) in AWGN channels. The greedy heuristic of the algorithm is to isolate portions of time series signals that are similar to a dictionary component with the highest inner product. Speech signals typically have a 12 dB PAPR for speech segment lasting several seconds [10]. Therefore, the first iterations of the algorithm will most likely detect valid speech components providing the dictionary elements are similar to time series

T. Bryan is a PhD candidate in the Electrical and Computer Engineering Department, Florida Institute of Technology, Melbourne, Fl, 32901 USA (phone: 321-626-3912; e-mail: jbryan@ my.fit.edu).

V. Kepuska is Associate Professor in the Electrical and Computer Engineering Department, Florida Institute of Technology, Melbourne, Fl 32901 USA (phone: 321- 674-7183; e-mail: vkepuska@ fit.edu

I. Kostnaic is Associate Professor with Electrical and Computer Engineering Department, Florida Institute of Technology, Melbourne, Fl 32901 USA (phone 321-674-7189 e-mail: kostanic@fit.edu).

portions of the data.

Lobo and Loizou implemented a matching pursuit atomic decomposition for voice activity detection using Gabor atoms [2]. In-class and out-of-class mixing matrices where used to reduce the dimensionality of the data. Classification was performed by a radial basis neural network that required supervised learning. Lobo and Loizou achieved 84% accuracy at a 5dB SNR from 64 randomly chosen TIMIT corpus sentences with AWGN [2].

This paper used 64 randomly chosen TIMIT corpus sentences with AWGN as a basis of comparison to the work of Lobo and Loizou using Gabor atoms. The gammatone atom is also used for atomic decomposition as it accurately models the auditory neural response in the human peripheral audio pathway to natural sounds as well as speech [3]-[6]. Kressner el al. used gammatone atomic decomposition implemented by matching pursuit to denoise speech. They demonstrated improved intelligibility in noisy environments for hearing aid applications [8]. Today gammagrams are used as an alternative to spectrograms to represent time-frequency characteristics of speech. Moreover, Gamma Frequency Cepstral Coefficients are being used in place of the FFT based Mel Frequency Cepstral Coefficients for feature vectors in speech processing. The VAD is implemented withGabor atoms, and then repeated using gammatone atoms is order to see if the gammatone atom VAD has better accuracy than the Gabor VAD in AWGN channel and at low SNR's.

The design parameters for gammatone and Gabor atoms are presented in Section II. Section III discusses the implementation of matching pursuit for the VAD. Finally, the results and conclusion are presented in Sections IV and V.

II. Gammatone and Gabor Atom Design Parameters

Gammatone filter banks were implemented using Infinite Impulse Response filters by Roy Patterson in 1992 as an efficient approximation to auditory frequency processing of the human ear [4]. The filter bank for this VAD uses logarithmically spaced center frequencies as opposed to the Equivalent Rectangular Bandwidth design proposed by Patterson. This is to enable comparison with equally space Gabor atoms.

There are 16 gammatone and 16 Gabor filters with a minimum frequency of 100Hz and a maximum center frequency of 3200Hz. The TIMIT data is resampled at F_s= 8ksps. The window length for all Gabor and gammatone atoms is 50mSec which corresponds to 400 samples at 8ksps. The logarithmically spaced center frequencies ($\boldsymbol{F_c}$) are given as:

$\boldsymbol{F_c}$ = [100, 126, 159, 200, 252, 317, 400, 504, 635, 800, 1008, 1270, 1600, 2016, 2540, 3200] Hz.

The gammatone impulse response is given by,

$$\Gamma(t) = At^{(n-1)}e^{-bt}cos(\omega_c t + \theta). \tag{1}$$

The pulse width is a function of the center frequencies($\boldsymbol{F_c}$),

$$b = B\boldsymbol{F_c}. \tag{2}$$

The value of A was set so that the L_2 norm = 1.

The value of θ was set so that the peak of the cosine function coincides with the peak of the gamma envelope.

The parameter n controls the rise time of the gamma pulse and was found to be 4 for human auditory response modeling [4]. The parameter B controls the bandwidth and was optimized for the best frequency overlap characteristic with logarithmically spaced center frequencies.

Gabor atoms used the same center frequencies $\boldsymbol{F_c}$, for purposes of comparison to gammatone atoms. The equation for the Gabor atom is given by,

$$\gamma(t) = \frac{1}{\sqrt{2\pi}\sigma}e^{-t^2/\sigma^2}cos(\omega_c t + \theta). \tag{3}$$

The bandwidth of the Gabor atoms is set by σ, which increases logarithmically with center frequency. The vector $\boldsymbol{\sigma_{cs}}$was determined empirically, to match the frequency response of the gammatone atoms. The Gabor vector $\boldsymbol{\sigma_{cs}}$is,

$\boldsymbol{\sigma_{cs}}$ = [3.00, 3.70, 4.57, 5.65, 6.98, 8.61, 10.64, 13.13, 16.22, 20.03, 24.73, 30.54, 37.71, 46.56, 57.50, 71.00].

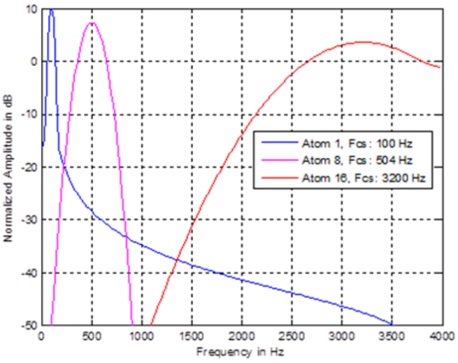

Fig. 1 Gabor L_2 Normalized Frequency Responses, atom1 = 100Hz, atom8 = 504Hz and atom16 3200Hz

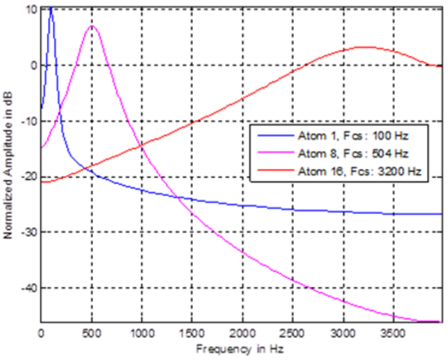

Fig. 2 Gammatone L_2 Normalized Frequency Response, atom1 = 100Hz, atom8 = 504Hz and atom16 3200Hz

The frequency responses of the Gabor and gammatone atoms 1, 8, and 16 are shown in Figs. 1 and 2. The normalized Gabor and gamma envelopes are shown in Figs. 3 and 4. The center frequencies for the Gabor and gammatone atoms match exactly, while the bandwidths are only approximated. However, one can see from Figs. 1 and 2, the bandwidths are very similar. Figs. 3 and 4 show the pulse widths in the time domain are also very similar between the Gabor atom and the Gammatone atoms. The Gabor atom has a sharper roll-off characteristic in the frequency domain. The Gammatone atom has a faster rise time in the time domain and a slow roll-off characteristic, while the Gabor atom is modulated by a symmetrical Gaussian pulse in the time domain.

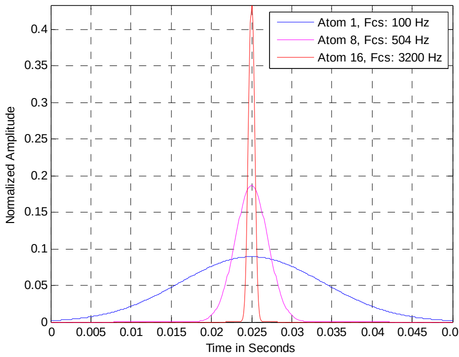

Fig. 3 Gabor L_2 Normalized Time Domain Envelopes, atom1 = 100Hz, atom8 = 504Hz and atom16 =3200Hz

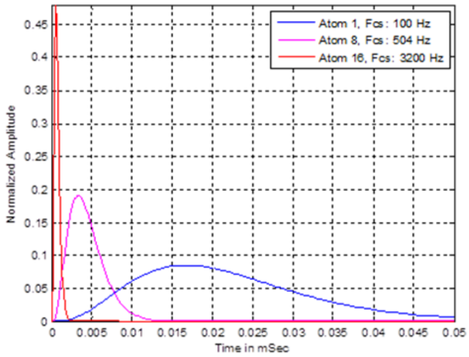

Fig. 4 Gammatone L_2 Normalized Time Domain Envelopes, atom1 = 100Hz, atom8 = 504Hz and atom16 =3200Hz

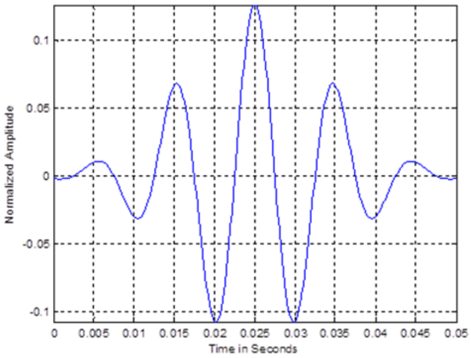

Fig. 5 Gabor L_2 Normalized Impulse Responses, atom1 = 100Hz

The Gabor and Gammatone impulse responses for atom1 which corresponds to the lowest frequency, 100 Hz is shown in Figs. 5 and 6. The highest frequency, 3200 Hz impulse responses for atom16 are shown in Figs. 7 and 8. The pulse widths are seen to be very similar for the lowest and highest frequency Gabor and Gammatone atoms. Additionally, the normalized amplitudes are also very similar.

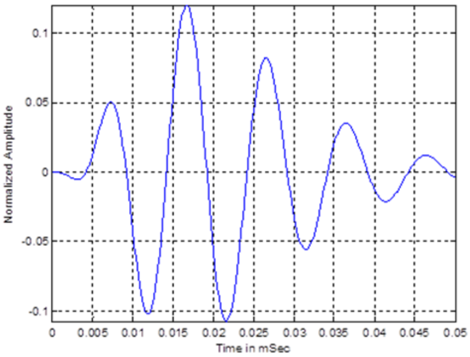

Fig. 6 Gammatone L_2 Normalized Impulse Responses, atom1 = 100Hz

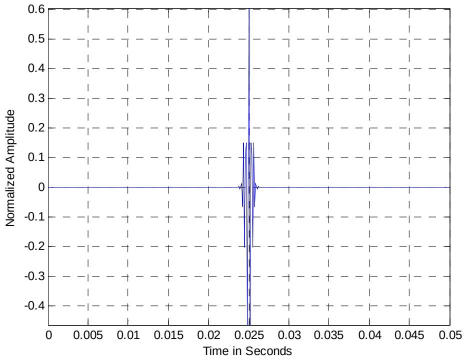

Fig. 7 Gabor L_2 Normalized Impulse Responses, atom16 = 3200Hz

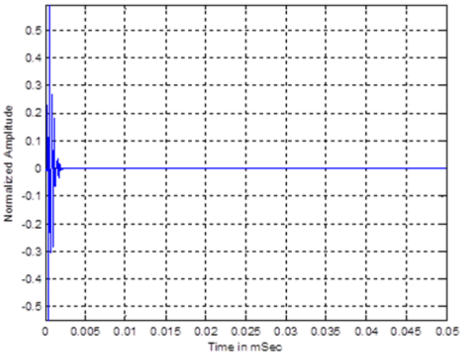

Fig. 8 Gammatone L2 Normalized Impulse Responses, atom16 = 3200Hz

III. Matching Pursuit VAD Implementation

Matching pursuit is a greedy algorithm that represents time series data as a linear superposition of fundamental atoms. Matching pursuit follows a simple heuristic of finding correlation peaks between input data and a set of atoms for all possible shifts of each atom. The algorithm finds correlation peaks that represent a minimum mean square error (MMSE) fit between portions of the data and the best atom provided the L_2 norm of the atom is set to 1. The algorithm halts once a stopping goal is achieved. If an amplitude coefficient falls below a minimum threshold, or when a certain portion of the energy of the signal is captured by the decomposition the algorithm may be halted. Alternatively, the algorithm might be terminated after a specified number of iterations. This VAD implementation uses a stopping goal that is specified by a fixed number of iterations based on the desired data compression.

The audio stream $x(t)$ may be represented by a linear superposition of atoms $\varphi_m$from a dictionary $\mathcal{D}$.

$$x(t) = \sum_{m=1}^{M} \sum_{i=1}^{n_m} s_{m,i} \varphi_m\left(t - \tau_{m,i}\right) + e(t) \tag{4}$$

where, $e(t)$ is the approximation error or residual, after n iterations of matching pursuit. The input data length is designated by $L,$ so that the number of iterations of the algorithm is given by,

$$n = DL. \tag{5}$$

where, D is the data compression. The optimum data compression for a given input SNR was found by computer simulation. The reconstructed SNR produced by matching pursuit was compared to the noiseless TIMIT input. The computer search was conducted using an iteration count ranging from 1% to 30% of the number of samples in the TIMIT sentence. This corresponds to a data compression of 99% to 70%. A typical TIMIT sentence lasts 3 to 5 seconds. Using an 8ksps sample rate this yields 24k to 40k samples per sentence. For data compression of 99%, or 1% of the sample length, typical iteration counts range from 240 to 400 iterations of matching pursuit.

Simulation parameters:

- The length of the input TIMIT sentence, L.
- Number is iterations, $n = DL$.
- The number of Gabor and gammatone atoms, M.
- The index of the correlation peak, i.
- The amplitude coefficient for the time index i and atom m is denoted by $s_{m,i}$.
- The time index i and atom m coefficient is denoted by $\tau_{m,i}$.

The steps of matching pursuit are

1. Initialize the algorithm $\mathcal{R}_0 = x(t)$
2. Compute for all $\varphi_{m,} \in \mathcal{D}$: CORR$(\mathcal{R}_{n-1}, \varphi_{m,})$ = $|\langle\mathcal{R}_{n-1}, \varphi_{m,}\rangle|$
3. Find the largest inner product, $\text{maxArg}(|\langle\mathcal{R}_{n-1}, \varphi_{m,}\rangle|)$
4. Compute the new residual, $\mathcal{R}_n = \mathcal{R}_{n-1} - \langle\mathcal{R}_{n-1}, \varphi_{m,}\rangle\varphi_{m,}$
5. Repeat step 2-4 until n iterations of the algorithm are complete.

The matching pursuit algorithm coherently finds the best match between the data and the atoms of the dictionary$\mathcal{D}$. The correlation process phase aligns the best atom with input data while at the same time minimizing the MMSE between the atom and the data segment. The simple heuristic of selecting the largest correlation peak from the best atom in the dictionary effectively produces a time overlapped MMSE detection of the data, which is known to optimize the mutual information between the data and the dictionary elements [9]. This approach works well so long as the atoms accurately represent the data. For this reason gammatone and Gabor atoms are compared for their performance in representing speech.

Examples of speech segments that are similar to gammatone and Gabor atoms are shown in Figs. 9 and 10. Both the Gabor and gammatone atoms have the same correlation index for the first iteration of matching pursuit. It can be seen that neither atom has a perfect fit to the data. This is primarily due to poor frequency resolution associated with using only 16 atoms for the VAD implementation.

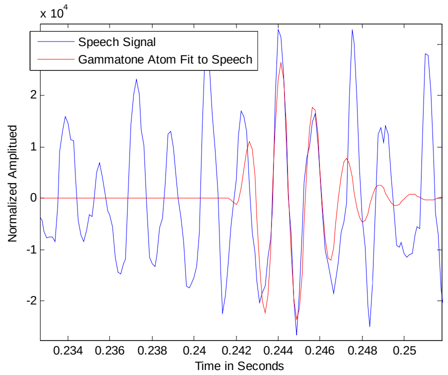

Fig. 9 Structure of Speech Similar to a Gammatone atom

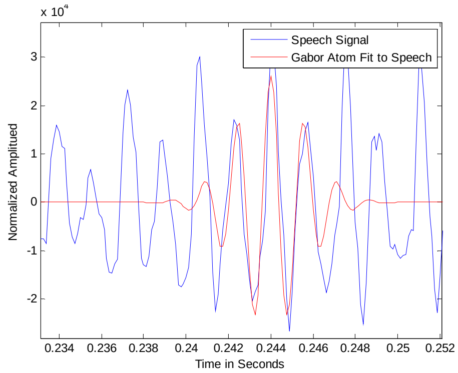

Fig. 10 Structure of Speech Similar to aGabor atom

The denoising capability of matching pursuit for high noise environments can be seen in Figs. 11 and 12. At a -5dB input SNR, the output achieves a peak of 1.5dB for 99.2% data compression. This shows an improvement of 6.5 dB from the input to output SNR. For a -2.5dB input SNR, the peak is at 99% data compression with an SNR improvement of 5dB.

Fig. 13 shows a peak of 2.8dB for 98% data compression at 0 dB input SNR. This is an improvement of 2.7 dB from the input to output SNR. For a 5dB input SNR, the peak is at 96.3% data compression. For 5 dB SNR's and higher, matching pursuit no longer is able to perform denoising as can be seen in Fig. 14. Fig. 15 shows a 10 dB input SNR has a reduction of 3.5 dB SNR at 92% data compression. Fig. 16

shows a 20 dB input SNR exhibits a 10dB SNR loss at a data compression below 82%. This study was optimized for a 5 dB input SNR and therefore, a 96.3 % data compression goal was used as the stopping criteria for matching pursuit. This threshold was low enough to capture some of the low SNR performance, and high enough to still perform adequately for high SNR inputs.

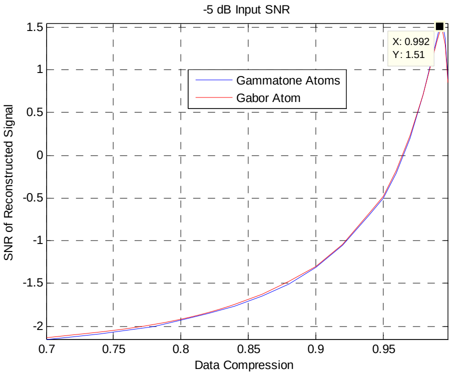

Fig. 11 Matching Pursuit Reconstruction SNR Verses Data Compression for a -5dB SNR

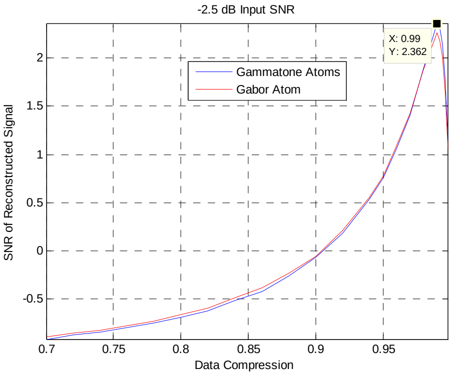

Fig. 12 Matching Pursuit Reconstruction SNR Verses Data Compression for a -2.5dB SNR

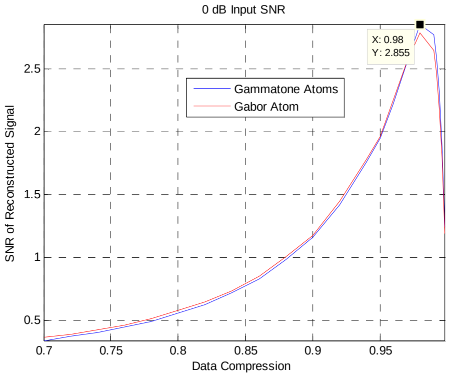

Fig. 13 Matching Pursuit Reconstruction SNR Verses Data Compression for a 0 dB SNR

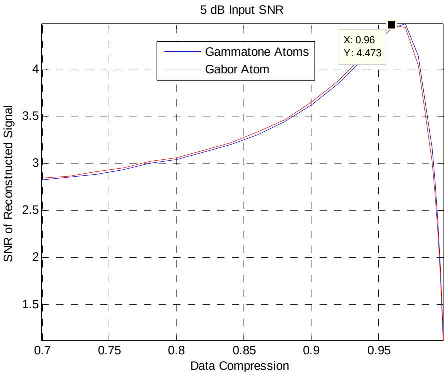

Fig. 14 Matching Pursuit Reconstruction SNR Verses Data Compression for a 5dB SNR

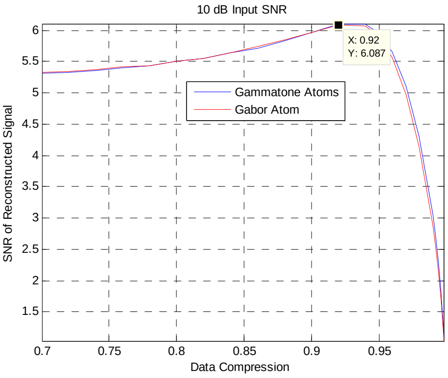

Fig. 15 Matching Pursuit Reconstruction SNR Verses Data Compression for a 10 dB SNR

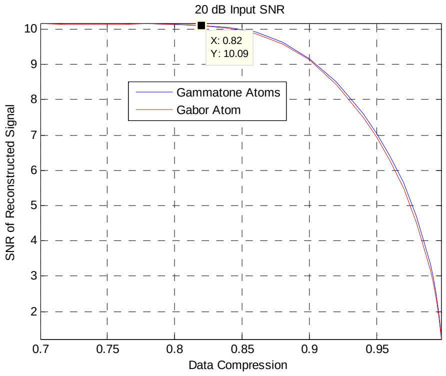

Fig. 16 Matching Pursuit Reconstruction SNR Verses Data Compression for a 20dB SNR

Matching pursuit was run on 64 randomly selected sentences from the TIMIT database. Due to the random nature of TIMIT sentence selection, 10 trials of 64 randomly selected TIMIT sentences were run to get a good mean statistic of the accuracy. The data for each sentence was collected from 0 to 20 dB SNR in 1 dB steps, with the number of iterations n set to 3.7% of the sample length. After n iterations of matching pursuit, the algorithm summed the amplitude coefficients, $s_{m,i}$for each atomic index$m$. The sums were sorted and only the best 50% of the atomic indices were used to reconstruct the signal.

$$A_{sum}(m) = \sum_{i=1}^{n_m} s_{m,i}\, for\ m = 1,2 \dots M. \quad (6)$$

$$m_{best} = sort\left(A_{sum}(m)\right)\ take\ top\ 50\%. \quad (7)$$

This produced a very sparse representation of the data that had 98.15% data compression.

The m_{best}atoms were used to reconstruct the data. The envelope of the reconstructed data was used for the VAD. An adaptive threshold, based on an estimated SNR, was applied to the envelope to produce the VAD output.

The high energy portions of the reconstructed signal were used estimate the signal strength. Due to the sparsity of the representation, much of reconstruction had zero energy, even at low SNR's. The noise floor was estimated by finding the mean of the L_2 norm of the noisy speech where the speech estimate was zero. The noise floor estimate was multiplied by a noise threshold factor that produces an envelope detection threshold. A noise floor multiplier was found by trial and error for input SNR's, from 0 dB, in steps of 2dB to 20 dB. The noise floor multiplier was manually adjusted to produce the best VAD accuracy for a particular input SNR. This data was paired with an SNR estimate to produce the following data vectors,

Estimate of SNR in dB = [-8.94, -4.59, 1.16, 6.33, 9.47, 11.58, 13.46, 15.08, 16.46, 19.25]

Optimum Noise Floor Multiplier = [20, 22, 27, 40, 53, 67, 80, 100, 117, 130]

Polynomial regression was used to generate a quadratic model of the SNR estimate that produces the noise multiplication threshold. The quadratic model takes the following form,

$$\text{Signal detection Threshold} = \boldsymbol{X\,\theta}$$

where $\boldsymbol{X}$ is a quadratic vector of the SNR estimate,

$$\boldsymbol{X} = [\ 1\ snrEst_{dB}\ snrEst_{dB}^2\ ],$$

with model parameters,

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$$

The model was found by regression by stacking the X's into rows and targets Y into a column

$$\boldsymbol{X} = \begin{bmatrix} 1\ snrEst_{0dB}\ snrEst_{0dB}^2 \\ 1\ snrEst_{2dB}\ snrEst_{2dB}^2 \\ 1\ snrEst_{4dB}\ snrEst_{4dB}^2 \\ \cdot \\ : \\ 1\ snrEst_{20dB}\ snrEst_{20dB}^2 \end{bmatrix}, \boldsymbol{Y} = \begin{bmatrix} 20 \\ 22 \\ 27 \\ \cdot \\ : \\ 130 \end{bmatrix}$$

Then the model was calculated by the normal equation as,

$$\boldsymbol{\theta} = (\boldsymbol{X^T}\mathbf{X})^{-1}\boldsymbol{X^T}\mathbf{Y}$$

For 64 randomly selected sentences from the TIMIT corpus, the model parameters were found to be,

$$\boldsymbol{\theta} = \begin{bmatrix} 21.57 \\ 1.74 \\ .21 \end{bmatrix}$$

Fig. 17 shows the model fit to the optimized data. Note that the SNR estimate ranges from -10 to 20 dB, while the actual input SNR was varied from 0 to 20 dB. The signal measurement has additive noise which corrupts the measurement, while the noise floor estimate contains low energy speech which also adds error. The error was of little consequence as the noise floor quadratic gain model maps the SNR estimate to the optimum noise floor gain. This produces the optimized VAD detection threshold and was very repeatable for 64 randomly selected TIMIT sentences.

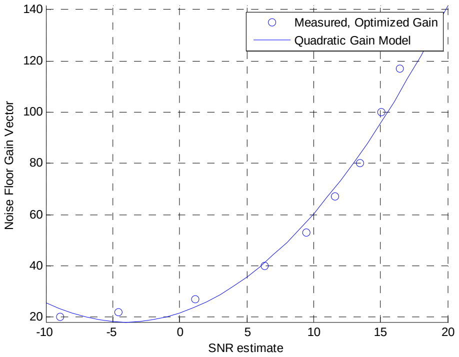

Fig. 17 Quadratic Model of Noise Floor Gain Optimized for VAD Accuracy

IV. Results

The VAD accuracy was measured by generating VAD time indices for a 30 dB reference input SNR. The AWGN was adjusted for desired input SNR and the VAD was rerun. The accuracy was based on the ratio of common to common plus different VAD time indices. Examples of the performance of the VAD for a reference SNR of 30 dB, and 5 dB SNR are shown in Figs. 18 and 19. The reconstructed envelope for 98.15% data compression, shown in magenta can be seen for both 30dB SNR's and a 5 dB SNR. The VAD is able to isolate portions of speech at low SNR that appear in the reference. The VAD output is shown in black and was 90% accurate to the VAD performance at 30 dB SNR for data collected from 64 random TIMIT sentences.

Trial runs of 10 different repetitions of 64 TIMIT random sentences were run to obtain a mean estimate of the VAD performance. The VAD accuracy verses input SNR can be seen in Figs. 20 and 21. The VAD achieves 70% accuracy at 0 dB SNR, 90% accuracy at 5 dB SNR and 97% accuracy at 20dB SNR. There was no distinguishable difference in the performance of Gabor verses gammatone atoms. The variance of the Gabor atoms appears slightly lower than the gammatone atoms. However, the mean performance looks nearly identical.

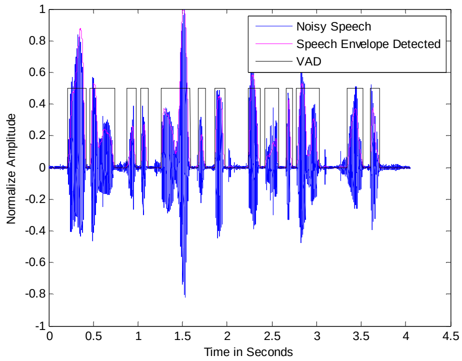

Fig. 18VAD Performance, 30 dB SNR

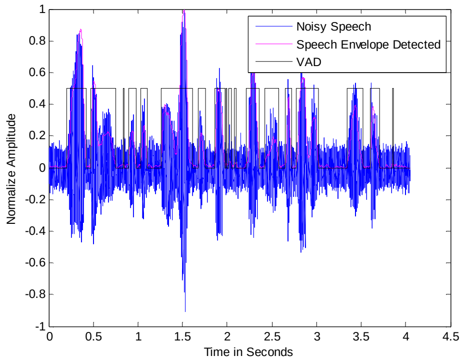

Fig. 19 VAD Performance, 5dB SNR

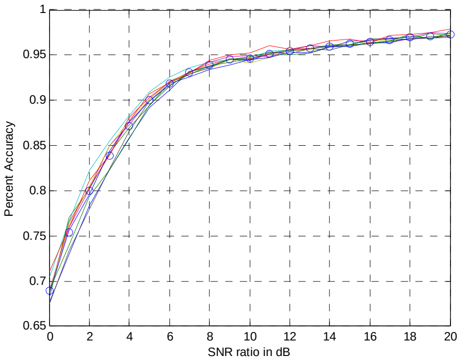

Fig. 20 VAD Accuracy, Gabor Atoms, 98.15% Data Compression, 64 Random TIMIT Sentences, 10 Trials

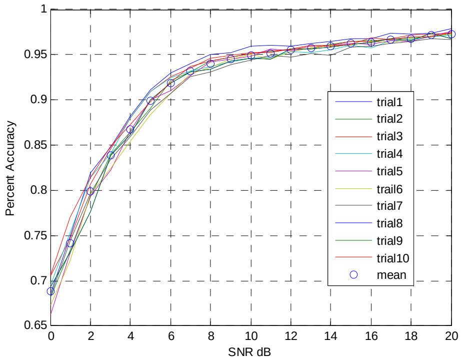

Fig. 21 VAD Accuracy, GammatoneAtoms, 98.15% Data Compression, 64 Random TIMIT Sentences, 10 Trials

V.Summary and Conclusions

Matching pursuit atomic decomposition for speech signals produces a high quality VAD with good accuracy at low SNR's. The stopping criterion for matching pursuit was based on a fixed data compression value. The optimum data compression of 96.3% for a 5 dB SNR was used for all input SNR's with good results. The high data compression allows matching pursuit to isolate high energy portions of speech at low SNR's. The intuition for this performance is that speech has a high PAPR whereas AWGN does not. Consequently the first iterations of the algorithm are most likely to find valid speech components. This approach should work well with any type of signals with high PAPR's in AWGN environments. The salient feature of matching pursuit is the atoms closely match the signature time-frequency characteristics of the data. An additional nuance of the VAD implementation was to retain only the atoms with the best fit to the data based on the atom index. This approach adapts to the time-frequency characteristic for individual speakers automatically. Finally, an adaptive reconstructed envelope threshold was based on a SNR estimate of the noisy speech. The signal estimate was based on the high energy portions of the data. The zero energy indices found by the sparse envelope reconstruction of matching pursuit were used to calculate a noise floor estimate of the data base on the average L_2 norm of the samples. The SNR is actually a signal plus noise to noise pulse signal estimate. The SNR estimate was mapped to optimal thresholds found by trial-and-error. A quadratic model was derived that maps the target thresholds to the SNR estimate for a range of input SNR's from 1 to 20 dB.

The simple adaptive VAD atomic decomposition has 90% accuracy as compared to the VAD of Lobo and Loizou which has 84% accuracy at a 5 dB input SNR. Both approaches use matching pursuit atomic decomposition with Gabor atoms, and both used 64 randomly selected sentences from the TIMIT corpus. The later approach however, is far more complicated as it requires supervised learning as opposed to this simple adaptive implementation

REFERENCES

[1] Gabor, D., Theory of communication, J. Inst. Elect. Eng., 93, pp. 429–457. 1946

[2] Lobo, A., Loizou, P., Voiced/unvoiced speech discrimination in noise using Gabor atomic decomposition. ICASSP (1) 2003: 820-823

[3] Smith, E., Lewicki, M., Efficient auditory coding. Nature, 439(7079):978–82, 2006.

[4] R. Patterson I. Nimmo-Smith. An Efficient Auditory Filterbank Based on the Gammatone Function. Institute of Acoustics on Auditory Modelling 1987

[5] Slaney, M., (1998) "Auditory Toolbox Version 2", Technical Report #1998-010, Interval Research Corporation, 1998.

[6] Atlas, L. Decomposition of speech and sound into Modulations and Carriers. http://msrvideo.vo.msecnd.net/rmcvideos/173320/dl/173320.pdf, Microsoft Research & University of Washington. 2012

[7] Mallat, S., Zhang, Z., Matching Pursuits with Time-Frequency Dictionaries. IEEE transactions on signal processing, Vol 41. No 12, 1993

[8] Kressner, A., Anderson, D., Rozell, C. Causal Binary Mask Estimation for Speech Enhancements using Sparsity Constraints. Proceedings on Meetings on Acoustics Vol. 9, 055037 2013

[9] Guo, D., Verdu', S., Mutual Information and Minimum Mean-Square Error in Gaussian Channels. IEEE transactions on information theory, Vol. 51, No. 4, 2005

[10] Eargle, J., Handbook of Recording Engineering. 4th Addition. Springer Science and Business Media. ISBN 1-4020-7230-9 (HC), 2003.