Low-Power OZGF Bank and MR Hamming Windowing for Embedded Speech Recognition

BRIAN SMITH
The Applied Research Laboratory
The Pennsylvania State University
University Park, Pennsylvania
USA
bms24@arl.psu.edu

JOHN SUSTERSIC
The Applied Research Laboratory
The Pennsylvania State University
University Park, Pennsylvania
USA
jps263@arl.psu.edu

MICHAEL MOORE
The Applied Research Laboratory
The Pennsylvania State University
University Park, Pennsylvania
USA
mjm83@arl.psu.edu

Abstract: We present novel implementations of a One-Zero Gammatone Filter and a multiresolution Hamming Window with constant time complexity for low power digital implementation in embedded speech recognition systems. We compare our model with state-of-the-art basilar membrane models in terms of computational complexity and in terms of phone classification accuracy on the TIMIT dataset and show quantitative advantages in both, enabling better speech recognition for a broader class of power and resource constrained digital embedded systems.

Key–Words: Gammatone Filterbank, Phone Recognition, Embedded Systems

1 Introduction

Decades of research in auditory filters have yielded many models that either strive to mimic the known response of human basilar membrane, to be used in phone detection/classification or other natural language processing, or frequently both guided by the intuition that the nonlinear filtering in the human ear in real ways directly supports auditory perception. The structure of the ear is such that different frequencies resonate a different points along the basilar membrane yielding a tonotopic organization which can be modeled by an array of overlapping band pass filters usually called ’auditory filters’ in this context [1]. Each auditory filter represents basilar membrane displacement at a point along the membrane which is generally non-linear and level dependent. Auditory filters decrease in bandwidth as their center frequency decreases [2, 3].

Classically, speech analysis research has centered strongly on (power) cepstral and mel-frequency cepstral analysis, where the power cepstrum is defined as follows [4]:

$$\left|\mathcal{F}^{-1}\left\{\log\left(\left|\mathcal{F}\left(f(t)\right)\right|^2\right)\right\}\right|^2 \tag{1}$$

making use of the Discrete Fourier Transform (DFT). The term Mel-Frequency Cepstral Coefficients (MFCCs) refers to the cepstrum whose frequency domain is first filtered according to a mel scale designed to better fit psychoacoustic data on pitch perception [5, 6]. Though MFCC analysis is pervasive in natural language processing, it does not strive necessarily to directly model basilar membrane displacement or its dynamic adaptability.

Gammatone (GT) filters have also been widely used to model auditory filters for natural language processing. A GT filter’s impulse response is a sinusoid with the amplitude envelope of a scaled gamma distribution [7]. Patterson’s Ear defines a parallel filter bank of GT filters with various parameters selected to match psychoacoustic data [7, 8, 9] which like mel scales results in non-uniform spacing of frequencies (though the frequencies are selected as a consequence of mimicking basilar membrane response instead of perceptual pitch relations). In this model, Glasberg-Moore’s suggested parameters are often selected since the asymptotic limits for the filters’ Q values are greater than other published parameters [7]. A broad comparative overview of auditory filter models finds that the One-Zero Gammatone Filter (OZGF) [10, 11] and the Pole-Zero Filter Cascade [10] are comparatively superior choices as auditory filters models among a pool of rounded exponential, GT and filter cascade models [10]. While GT filters are most frequently used in cochlear modeling and speech recognition, these filters are nearly symmetric in the pass band (where significant asymmetries exist in biological transfer functions) and it is difficult to parameterize the GT for level-dependent adaptations in the auditory filter [11].

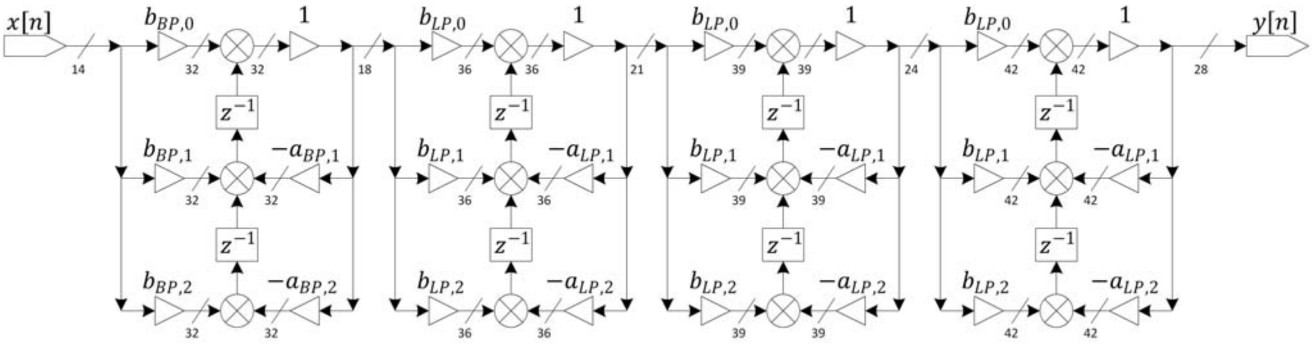

Figure 1: 4^{th} Order Digital Biquad with data paths sized to maintain stability of OZGF for full-range input.

2 One-Zero Gammatone Filter Bank and Multiresolution Hamming Windowing

The OZGF transfer function is defined as follows [11]:

$$H(s) = \frac{\omega_0^{2N-2}}{\left[s^2 + \frac{w_0}{Q}s + w_0\right]^{N+1}} \times \frac{w_0(s+a)}{s^2 + \frac{w_0}{Q}s + w_0} \quad (2)$$

where the real zero is chosen to be $a = w_0/Q$ [11]. Thus given the sampling frequency f_s, center frequency $f_c = w_0/2\pi$, and filter factor Q, one may use the bilinear transform to map into the z-plane for the band pass and low pass biquads, and this transfer function may be expressed simply as a fourth order, direct form II (transposed) digital biquad cascade (Fig. 1).

A parallel bank of OZGFs may then be used to model basilar membrane response. We use the Glasberg/Moore [7, 8, 9] method to determine center frequencies for the filter bank, computing filter parameters over a range of $Q = 2, 3, \ldots, 12$ and we verified filter behavior using Matlab. We then implemented the filters in Verilog, and using ModelSim we validated the fixed-point Verilog filter performance in both the time and frequency domains against expected responses using Matlab. Note that the Biquad cascade illustrated in Fig. 1 features incrementally sized data paths to maintain stability in the IIR filter across the full dynamic range of the input.

Psychophysical data indicate that spoken language recognition requires a temporal resolution on the order of 10 milliseconds; thus we follow conventional methods for natural language processing and use Hamming windows to decimate the data in time, reducing the sampling rate to 100 Hz. Let $N_s = \lfloor T_s F_s \rfloor$, where $T_s = 0.01$ seconds is the windowing period ($2T_s$ is the window size) and F_s is the sampling frequency. The Hamming window function is then

$$w(t) = \begin{cases} 0.08 + 0.46\cos\left(\frac{\pi t}{2N_s}\right), & t \in [-N_s, N_s] \\ 0, & otherwise \end{cases} \quad (3)$$

where $2N_s$ is the number of samples in the Hamming Window. Let $y(t)$ be the signal to be windowed. We then compute the windowed response (in decibels) with half-square rectification and log compression:

$$y_w(T) = 20\log \sum_{i=-N_s, i\in\mathbb{Z}}^{N_s} w(i)\,|y(T+i)| \quad (4)$$

The same window function may then be used to provide a second temporal resolution to allow estimates of the first and second derivatives of $y_w(T)$. Windowing at twice the rate yields three distinct subwindows $y_w(T)$, $y_{w,c}(T)$, and $y_{w,r}(T)$:

$$y_{w,l}(T) = 20\log \sum_{i=-N_s/2, i\in\mathbb{Z}}^{N_s/2} w(2i)\,|y(T+i-N_s/2)| \quad (5)$$

$$y_{w,c}(T) = 20\log \sum_{i=-N_s/2, i\in\mathbb{Z}}^{N_s/2} w(2i)\,|y(T+i)| \quad (6)$$

$$y_{w,c}(T) = 20\log \sum_{i=-N_s/2, i\in\mathbb{Z}}^{N_s/2} w(2i)\,|y(T+i+N_s/2)| \quad (7)$$

Thus $y_{w,r} := (T)y_{w,l}(T+1)$. Maximum likelihood estimates of first and second derivatives with respect to time may then be defined as follows:

$$\hat{y}'_w(T) = \frac{y_{w,r}(T) - y_{w,l}(T)}{T_s} \tag{8}$$

$$\hat{y}''_w(T) = \frac{y_{w,r}(T) - 2y_{w,c}(T) + y_{w,l}(T)}{T_s} \tag{9}$$

As with digital filters, computational complexity of the windowed data and estimates of the first two time derivatives is constant for each input sample and thus easily implemented in real-time. The OZGF is advantageous over other basilar membrane models because of its unique numerical advantages. Specifically, gammatone filters are eighth-order IIR filters, and like all high-order filters of this class are highly susceptible to instability and to quantization of their coefficients[12]. Coefficients for the Moore model gammatone filter with a center frequency of 100 Hz range from -54.2438 to 67.1041 (feedback) and from -8.82×10^{-8} to 2.12×10^{-8} (feedforward). Thus in this filter of this center frequency, coefficients range over ten orders of magnitude. At a center frequency of 10KHz, feedforward coefficients range from 0.14×10^{-2} to 3.22×10^{-2}. Floating point arithmetic of double precision is required to avoid instability in the filter for a filter bank spanning the frequency range of 100 Hz to 10KHz. In contrast, the FPGA design of our OZGT filter as a digital biquad allows us to implement the gammatone filter completely in fixed point arithmetic. For comparison, a double precision floating point multiplier implemented using Altera IP on Stratix IV devices requires ten 18-bit DSP resources along with several hundreds of other resources and a minimum of five clock cycle latency, and a single fixed-point 54-bit multiplier requires on three DSP resources and provides a single clock latency[13]. Thus to achieve the same rate of computation, floating point requires a clock at least five times that of fixed point and significantly more on-chip resources; since power required is proportional to both number of transistors in the circuit and to clock speed, floating point is an order of magnitude more expensive simply in terms of DSP resources and clock speeds, and even more so considering the additional FPGA resources required to implement floating point arithmetic. Thus the OZGT filter implementation significantly more efficient not only in terms of computational efficiency, but also in terms of power and resource requirements, and we have shown the filters are stable across the range of frequencies from 100 Hz to 10KHz.

An example of an audio recording of a male speaker uttering the phrase “the quick brown fox jumps over the lazy dog” after filtering by a OZGF bank ($Q = 10$) and windowed using these methods is illustrated in Fig. 2.

3 Methodology

To determine the appropriateness of applying the OZGF for embedded acoustic classification, we chose the task of phone identification using the standard TIMIT dataset [14]. We follow a similar method as [15]. First, unsupervised learning techniques are used to learn and select appropriate feature vectors by analysis of an unlabeled dataset. Then, these feature vectors are used to train a support vector machine (SVM) using the method of [16]. We report phone classification results using feature vectors learned from the OZGF filtered data with various values of Q, from the Moore-Glasberg model of the gammatone filter, and from the short-term Fourier transform (STFT) frequency response (at both uniformly sampled frequencies and at mel-frequency spacings).

The best performance on the TIMIT phone identification task has recently been reported in [15], using one of a set of unsupervised learning methods which have recently gained significant attention for success in classification in a variety of fields including handwritten digit identification[17] and computer vision[18], as well as phone classification. In these application it is often time-consuming and costly to hand-design the features upon which an algorithm will base its decisions. Very briefly, the technique consists of treating the spectrogram (short-term Fourier transform, or STFT) of an audio signal as an image, and then identifying phones in that image using features learned from an unlabeled dataset. We compare the performance of the OZGF by using images constructed from gammatone filtering, rather than the spectrogram images from the STFT.

To compute features to be used in the classification task, we construct and train a sparse autoencoding neural network (SAENN)[19]. A SAENN, a form of manifold learning, is trained to learn a sparse representation of a dataset, tailored to the class of data under investigation, and the basis of this representation become features for classification. Specifically, our neural network is a function

$$\hat{x} = f\left(W_T^2 \times f\left(W_1 x + b_1\right) + b_2\right) \tag{10}$$

where x and $\hat{x}$ are $n_v \times 1$ data vectors, W_1 and W_2 are $n_h \times n_v$ weight matrices, b_1 and b_2 are $n_v \times 1$ and $n_h \times 1$, respectably, sized bias vectors, and $f()$ is a the vectorized sigmoid function. We train the network by minimizing the cost

$$J(W,b) = \frac{1}{2m}\sum_{i=1}^{m} ||\hat{x}_i - x_i||_2 + R(W_1, W_2) + D(\rho||\hat{\rho}) \tag{11}$$

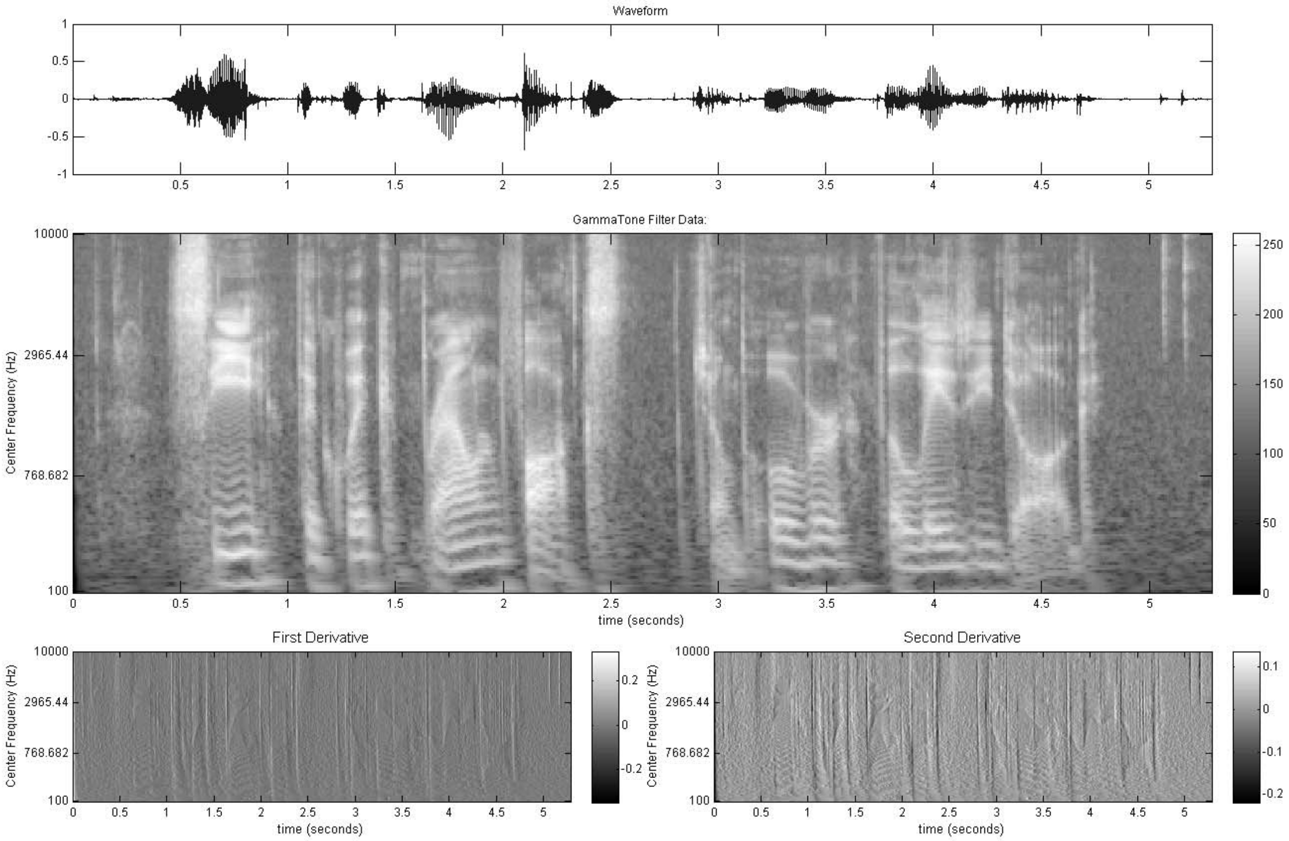

Figure 2: Filtered and windowed example of spoken English using 128 channel OZGF bank, $f_{min} = 100Hz$, $f_{max} = 10kHz$, $Q = 10$.

using a gradient descent optimization on W_1, W_2, b_1 and b_2. In the above cost function, m is the number of training examples, R is a regularization cost and D is a cost to promote sparsity so that each neuron in the hidden layer is active for a desired ρ fraction of the training examples

$$R(W_1, W_2) = \frac{\lambda}{2}\left[Tr(W_1W_1^T) + Tr(W_2W_2^T)\right] \tag{12}$$

$$D(\rho||\hat{\rho}) = \beta\sum_{j=1}^{n_h}\rho\log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j} \tag{13}$$

where β and λ are tunable weight parameters and $\hat{\rho}_j$ is the average activation of the jth hidden node. Once the cost function is minimized, the n_h rows of W_1, each of length n_v, are taken as feature vectors. To determine proper feature vectors for audio, we first applied the OZGT filter with $Q = 10$ to the training examples of the TIMIT data set, using 120 frequency components spaced logarithmically between 100 and 7500Hz, and Hamming window filtered the output to 20ms length segments with 10ms overlap. We similarly did so using the Glasberg/Moore model, various other values of Q for the OZGT filter, a spectrogram using the FFT with 128 uniformly spaced frequency components, and a spectrogram using the discrete Fourier transform (DFT) at the same frequency sample spacings as the gammatone filters, for fair comparison. From each sound file in the training set of the TIMIT database, two “patches” were randomly selected as training examples: a patch consists of the log-power of the 120 (or 128) frequency components taken over 4 consecutive Hamming window samples, for a width of 50 ms. Thus, the input to the autoencoding neural network consists of $m = 9240$ data vectors, of length $n_v = 480$ (or 512). Other parameters that were used (after selection via some empirical observation for tuning) are $\rho = 0.08$, $n_h = 300$, $\beta = 0.01$, and $\lambda = 0.01$. Gradient descent was used to learn (locally) optimal W_1, W_2, b_1 and b_2.

Table 1: TIMIT Phone Classification Accuracy: Correct phone identification

Method	Correctness
OZGF, Q=2	58.4%
OZGF, Q=4	60.6%
OZGF, Q=6	60.8%
OZGF, Q=8	60.9%
OZGF, Q=10	60.8%
OZGF, Q=12	60.6%
MGGF	61.2%
Fourier-Equal Frequency Spacing	52.4%
Fourier-Log Frequency Spacing	45.4%

For training the support vector machine classifier, feature vectors were built in a similar method to [16]: For each labeled phone training example, the feature vector consists of the activations of the hidden layer of the neural net from the "patch" at the beginning of the phone, concatenated with the activations from the patch at the end of the phone, concatenated with the average activation taken over all "patches" contained within the phone. The length of the phone was taken as an additional feature, meaning each SVM input consists of a 901 length vector. One-vs-all classification was performed, and results are reported on the grouped-phone (26-class) classification for the testing subset of the TIMIT database. This methodology is conceptually similar to the unsupervised feature extraction and phone identification work from [15], which, when including MFCC's in their feature vector as well, provided the best known performance.

4 Results

This investigation was performed to determine whether the features learned from using a OZGF performed at a similar level of performance to those derived from the more complex Moore/Glasberg Gammatone Filter model (MGGF), and to fairly compare their performance with a standard Fourier transform derived model. Results for the various filter models are reported in Table I.

For each filter, four sets of basis function feature vectors were derived by the neural network from random weight initializations, and the average performance of the classifications are reported. The typical variation between two trials of the same filter type was between 0.1% and 0.2%. We can see that for values of Q between 8 and 10, the classification performance is the highest, comparable with and even exceeding the performance of the Moore/Glasberg model. Performance also exceeds that of the simple uniformly spaced FFT and logarithmically spaced DFT log-power spectrograms. Additionally, we have broken out classification performance via each of six phone classes, as defined in [20]. It was seen that the performance of the classifier on fricatives peaks earlier with respect to Q than other classes (the performance of stops improves with greater Q). This implies that multiple filters may be used in parallel to achieve the greatest performance.

The best direct performance comparisons with our method are can be found in the analogous experiments from [15]. In that work the classification results (without the addition of MFCCs) use restricted Boltzmann models, rather than the autoencoding network model, and do not use biologically inspired filters. That experiment achieves comparable results of between 56.7% and 64.4% classification accuracy. The paper also demonstrates the fusion of unsupervised learning methods for feature extraction (explored in our work) with traditional mel-frequency cepstral coefficients, and achieves the best so-far reported performance for recognition on the TIMIT dataset, of 80.3%. We thus propose the fusion of OZGF features with MFCC features for future work, where our classification results could better approach state-of-the-art (but would also lose the computational efficiencies gained from discarding the Fourier transform model).

5 Conclusions

From this work, we can see that the OZGF provides computational advantages over the Moore/Glasberg gammatone filter when implemented in hardware, as well as classification performance advantages over the simple Fourier transform model. While filterbank models are not generally a feature of traditional automatic speech recognition systems (because the various components are strongly correlated and thus not well suited for the Gaussian Mixture Model paradigm in common use), deep belief networks are currently enabling their reintroduction. We propose using belief network structures in tandem with the one-zero gammatone filter model, rather than traditional Fourier transform filters, to increase recognition capabilities and to increase the ease of efficient implementation in hardware. While our classification results are not yet state of the art, this experiment demonstrates the improvement gained by using the OZGF over the Fourier transform. Future work includes augmenting the research of [15, 21] by including the one-zero gammatone filter, to attempt to improve on the state of the art classification algorithms.

Acknowledgements: This material is based upon work supported in part by the Office of Naval Research under grant number N00014-11-1-0893. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the sponsor.

References:

[1] Munkong, R.: Auditory Perception and cognition. IEEE Signal Processing Magazine, Vol 25 No 3, pp. 98-117 (2008)

[2] Gelfand, S. A.: Hearing: an introduction to psychological and physiological acoustics (4th ed.). Macel Dekker, New York (2007)

[3] Moore, B. C. J.: Parallels between frequency selectivity measured psychophysically and in cochlear mechanics. Scand. Audio Suppl. Vol 25, pp 129-52 (1986)

[4] Norton, M., Karczub, D.: Fundamentals of Noise and Vibration Analysis for Engineers. Cambridge University Press, Cambridge (2003)

[5] Harrington, J., Cassidy, S.: Techniques in speech acoustics. Springer. p. 18. (1999)

[6] Umesh, S., Cohen, L., Nelson, D.: Fitting the mel scale. In: Proc. ICASSP 1999 pp. 217-220 (1999)

[7] Slaney, M.: An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank. Apple Computer Technical Report #35 (1993)

[8] Glasberg, B. R., Moore, B. C. J.: Derivation of auditory filter shapes from notched-noise data. Hear. Res. Vol 47, pp. 103-138 (1990)

[9] Moore, B. C. J., Glasberg, B. R.: Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns. Hearing Research Vol 28 Nos 2-3, pp. 209-225 (1987)

[10] Lyon, R. F., Katsiamis, A. G., Drakakis, E.M.: History and future of auditory filter models. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems (2010)

[11] Katsiamis, A. G., Drakakis, E. M., Lyon, R.F.: Introducing the differentiated all-pole and one-zero Gammatone filter responses and their analog VLSI log-domain implementation. In: 9th IEEE International Midwest Symposium on Circuits and Systems, Vol 1 (2006)

[12] Kuo, Sen M., Lee, Bob H., Tian, W.: Real-time Digital Signal Processing: Implementations and Applications, 2nd edition, Wiley & Sons Ltd (2006)

[13] Altera, Inc.: Floating-point megafunctions user guide. `http://www.altera.com/literature/ug/ug_altfp_mfug.pdf` (2013)

[14] Fisher, W., Doddington, G., Goudie-Marshall, K.: The darpa speech recognition research database: Specifications and status. In: DARPA Speech Recognition Workshop (1986)

[15] Lee, H., Largman, Y., Pham, P., Ng, A. Y.: Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Neural Information Processing Systems (2009)

[16] P. Clarkson, P. J. Moreno: On the use of support vector machines for phonetic classification. In: IEEE Conference on Speech and Signal Processing, pp. 585-588 (1999)

[17] Ranzato, M. A.: Sparse feature learning for deep belief networks. In: Advances in neural information processing systems, Vol 20, pp. 1-8 (2007)

[18] Le, Q. V., Monga, R., Devin, M., Corrado, G., Chen, K., Marc'Aurelio Ranzato, J. D., Ng, A. Y.: Building high-level features using large scale unsupervised learning. In: Proceedings of the Twenty-Ninth International Conference on Machine Learning (2012)

[19] Goodfellow, I., Le, Q., Saxe, A., Lee, H., Ng, A. Y.: Measuring invariances in deep networks. In: Advances in neural information processing systems, Vol 22, pp. 646-654 (2009)

[20] Lopes, C., Perdigo, F.: Phone Recognition on the TIMIT Database. Speech Technologies pp. 285-302 (2011)

[21] Mohamed, A., Dahl, G. E., Hinton, G.: Acoustic modeling using deep belief networks. IEEE Transactions on Audio, Speech, and Language Processing, Vol 20 No 1, pp. 14-22 (2012)