A Machine Hearing System for Binaural Sound Localization based on Instantaneous Correlation

Ying Xu, Saeed Afshar, Ram Kuber Singh, Tara Julia Hamilton, Runchun Wang, André van Schaik
The MARCS Institute, Western Sydney University, Kingswood, NSW 2751, Australia
ying.xu@westernsydney.edu.au

***Abstract*— We propose a biologically inspired binaural sound localization system for reverberant environments. It uses two 100-channel cochlear models to analyze binaural signals, and each channel of the left cochlea is compared with each channel of the right cochlea in parallel to generate a 2-D instantaneous correlation matrix (correlogram). The correlogram encodes both binaural cues and spectral information in a unified framework. A sound onset detector is used to generate the correlogram only during the sound onsets, and the onset correlogram is analyzed using a linear regression approach as well as an extreme learning machine (ELM). The proposed system is evaluated using experimental data in reverberation environments, and we obtained an average absolute error of 16.5° for linear regression and 12.8° for ELM regression in the -90° to 90° range.**

Keywords*— *Electronic cochlea; Machine hearing; ELM; Neuromorphic engineering; Sound localization; Onset detection

I. INTRODUCTION

Sound source localization in reverberant environments is routinely performed by the humans but remains challenging for machine hearing systems. In human hearing, frequency-dependent binaural cues, such as interaural time difference (ITD) and interaural level difference (ILD), are exploited to determine the azimuthal angle of a sound in the horizontal plane. Due to the distance between the two ears, at lower frequencies, the ITD cue appears in the form of interaural phase difference (IPD), whereas at higher frequencies, the ITD shows in the form of interaural envelope delay (IED) [1]. In past decades, efforts have been made to implement a binaural sound localization system that can mimic the robust sound localization ability of the human auditory system. For example, van Schaik and Shamma implemented a neuromorphic sound localizer to determine the direction of incoming sound using two silicon cochleae [2], Chan et al. proposed a robotic sound localization system using a winner-take-all network to estimate the direction of the sound source through the ITD cue from a cochlea pair [3], Umbarkar et al. used a wave counting approach to estimate the sound source location [4], and Jin et al. implemented a sound localization system using the generalized cross-correlation (GCC) approach [5]. Additionally, computational models and algorithms have been proposed and developed for binaural localization. For example, May et al. exploited both the ITD and the ILD cues to train azimuth-dependent Gaussian mixture models (GMM) [6], Nix and Hohman combined the empirical statistics of the IPD with the ILD and proposed a Bayesian maximum a posteriori approach to estimate the location of a sound source [7], and Ma et al. used the cross-correlation function (CCF) and the ILD from each channel of a binaural model to train a separate deep neural network (DNN) [8].

Inspired by these studies, in this paper, we propose a binaural sound localization system based on the stereausis representation of binaural signals [9]. It makes use of two cochlear models to decompose the sound into frequency components and instantaneous correlation units to integrate the frequency dependent binaural cues from each frequency channel into a 2-D correlogram. An ELM is trained to learn the correlogram pattern during the signal onset to identify the azimuthal angle of the sound source. We have previously implemented the cochlear model and the ELM system on FPGA in [10] and [11] respectively, and these can be used together to create the proposed system.

II. SYSTEM ARCHITECTURE

A. CAR-FAC cochlear model

The proposed binaural sound localization system is shown in Fig.1. Two Cascade of Asymmetric Resonators with Fast-Acting Compression (CAR-FAC) cochlear models are used to analyze binaural signals. The CAR-FAC model is a biologically based digital cascaded auditory filter model that is proposed by Lyon in [1] and implemented on FPGA in [10]. It closely mimics the physiological elements that consist of the human cochlea. The CAR part models the basilar membrane (BM) function using two-pole-two-zero resonators. The FAC part includes a digital outer hair cell (OHC) model, a digital inner hair cell (IHC) model, and a spatial-temporal filter-loop. The architecture of the CAR-FAC model is shown in Fig.1(A). The CAR-FAC model shows an outstanding agreement with the biological data recordings comparing with other cochlear models and an improved signal-to-noise ratio (SNR) in each frequency channel [12]. Additionally, the CAR-FAC model shows a fine spectral representation of the sound source, and the IHC model provides more accurate IHC function than a simple half wave rectifier, which is beneficial to the pattern of the correlogram. Therefore, in the proposed system, we use two 100-channel CAR-FAC models at 32 kHz sampling rate with center frequencies (CF) ranging from 45 Hz to 2000 Hz according to the Greenwood mapping function [13]. The CAR-FAC FPGA implementation only takes 18% of the adaptive logic module (ALM), 24% of the memory blocks and 33% of the DSP blocks on a Cyclone V FPGA board. It is thus able to be integrated with other blocks to build a machine hearing system.

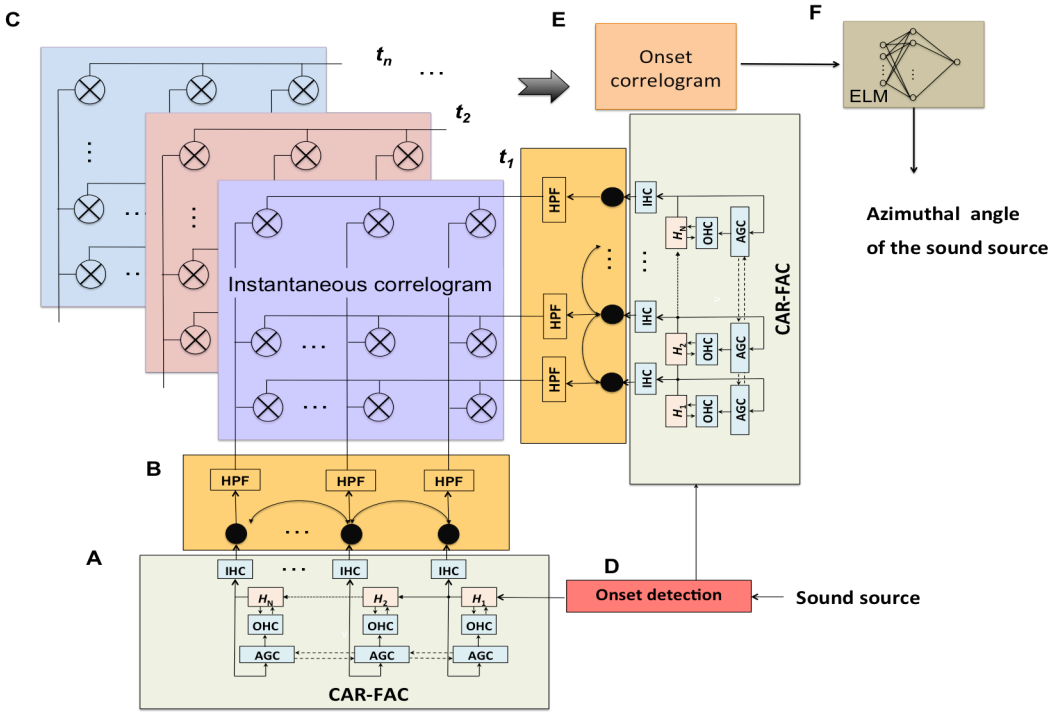

Fig. 1. Architecture of the binaural sound localization system. (A) The CAR-FAC model, H_1 to H_N are the transfer functions of the CAR part, the center frequencies (CFs) of the resonators decrease from H_1 to H_N. The OHC, the IHC and the automatic gain control (AGC) comprise the FAC part. The IHC output is connected to the lateral inhibition and a high-pass filtering (HPF) unit. (B) The lateral inhibition and HPF unit; It is consist of a high-pass spatial filter and a high-pass temporal filter. (C) The instantaneous correlogram, the circle marked with a cross is an instantaneous correlation unit, and the output from all the units forms a 2-D correlogram. (D) The onset detection function; it detects the onset of a sound source. (E) The onset correlogram; a short period after the onset detection, t_1 to t_n, includes n instantaneous correlograms. The n instantaneous correlograms are averaged to form the onset correlogram. (F) The ELM system; the onset correlogram is used to train the ELM to learn the azimuthal angle of the sound source.

B. Lateral inhibition and HPF

The IHC output from the CAR-FAC model is then connected to a lateral inhibition and a high-pass filtering (HPF) unit, as shown in Fig.1(B). The lateral inhibition mimics the spectral sharpening effects found in the cochlear nucleus of the auditory system and is modeled in [4] to enhance the representation of the harmonics for pitch detection. Additionally, we found that additional high-pass filtering improves the correlogram representation.

The lateral inhibition is modeled by a three-tap spatial filter across the cochlear channels:

$$L(i,t) = a\times IHC(i,t) + b\times IHC(i+1,t) + c\times IHC(i-1,t) \quad (1)$$

where i is the cochlear channel number, t is the discrete time, and $L(i,t)$ is the lateral inhibition output. a, b, and c are the spatial filter coefficients. Here a is set to 1.66, and b and c are set to -0.33. It is a three-tap sharpening filter with the coefficients summing to 1.

The HPF is connected to each cochlear channel after the lateral inhibition, and the cut-off frequency is set to be half of the CF of each frequency channel. The HPF output, $z(i,t)$, is given by:

$$z(i,t) = HPF(L(i,t)) \quad (2)$$

The spatial filter and the HPF can be straightforwardly implemented on FPGA with fixed-point numbers using multipliers and adders.

C. Instantaneous correlation

Instantaneous correlation units compute the correlations of the instantaneous activity of the two CAR-FACs output on each channel [14]:

$$Corr_{i,j}(i,j,t) = \begin{cases} 1, & if\ z_l(i,t)\times z_r(j,t) > 0 \\ -1, & if\ z_l(i,t)\times z_r(j,t) < 0 \\ 0, & if\ z_l(i,t)\times z_r(j,t) = 0 \end{cases} \quad (3)$$

where $Corr_{i,j}(i,j,t)$ is the instantaneous correlation from channel i of the left cochlea HPF $z_l(i,t)$ and channel j of the right cochlea HPF $z_r(j,t)$ at time t. Correlations of the same polarity of the two inputs produce a positive correlation signal, and correlations of the opposite polarity of the two inputs produce a negative anti-correlation signal. At time t, correlations of all the channels of the two CAR-FAC models comprise a 2-D instantaneous correlogram, as shown in Fig.1(C). The instantaneous correlation unit can be simplified and implemented on FPGA with fixed-point numbers by using comparators.

D. Onset detection

In the human auditory system, binaural cues are analyzed in the brainstem, and the direction of a sound source in the horizontal plane is identified shortly after the signal onset [15]. Additionally, in reverberant environments, signals inevitably include echoes after the signal onset, which greatly affects the sound localization system performance. Therefore, we propose a sound onset detection approach that detects the onset of the sound source to generate the correlogram during the signal onset to decrease the influence of echoes. The onset detection approach is given by:

$$\Delta E(t) = \log 10\left(\frac{\sum_{n=t}^{n=t+step} v(n)^2}{\sum_{n=t-step}^{n=t} v(n)^2}\right) \quad (4)$$

where $v(n)^2$ is the energy of the sound signal at time n, and $step$ is a time window. Here a threshold ΔEth is set, and $\Delta E(t)$ is compared with ΔEth. If $\Delta E(t) \geq \Delta Eth$ at time t, the onset time t is detected. The log10 function can be efficiently implemented on FPGA using a lookup table (LUT) [16], and equation (4) can be implemented on FPGA with fixed-point numbers by using two memory blocks, multipliers, adders, and a subtractor.

E. Onset correlogram

When an onset is detected, the onset correlogram is generated by:

$$Corr_{onset}(i,j) = \frac{1}{\Delta t \times f_s} \times \sum_{t=onset}^{t=onset+\Delta t} Corr(i,j,t) \quad (5)$$

where f_s is the sampling frequency and is set to 32 kHz. $onset$ is the detected onset time, Δt is a short period after the signal onset, $Corr(i,j,t)$ is the instantaneous correlation at time t, and $Corr_{onset}(i,j)$ is the averaged instantaneous correlograms during Δt. In the proposed system, the onset correlogram is transformed to have zero mean:

$$Corr_{elm}(i,j) = Corr_{onset}(i,j) - \frac{1}{100^2} \times \sum_{i=1}^{i=100} \sum_{j=1}^{j=100} Corr_{onset}(i,j) \quad (6)$$

where the zero-mean distributed $Corr_{elm}(i,j)$ is transformed from $Corr_{onset}(i,j)$ by subtracting its mean. Equation (5) and (6) can be implemented on FPGA with fixed-point numbers using memory blocks, multipliers adders, and subtractors.

F. ELM

The zero-mean distributed correlogram is then used to train the ELM to learn the azimuthal angle of the sound source. The ELM is able to produce good generalization performance and faster learning speed than the back-propagation algorithm [17]. Additionally, an online learning ELM algorithm, the Online Pseudo Inverse Update Method (OPIUM), has been proposed in [18], implemented on FPGA in [11], and demonstrated on the MNIST digits classification task. The system uses a time-multiplexing approach to reuse one module multiple times to build 8192 hidden layer neurons, which only takes less than 5% of the ALM, 1.4% of the memory blocks on a Cyclone V FPGA board [11]. It is thus able to be extended to a large network with more hidden neurons and integrated with the CAR-FAC module on FPGA.

III. Experiment and Evaluation

The proposed system is evaluated using experimental data from a reverberant environment (an office). The system is simulated with floating point numbers, and it can be transformed into the fixed-point numbers via well-established techniques [11]. In the experiment, the azimuthal angle ranging from -90 ° to 90 ° is divided into 13 locations with a 15 ° step. Two microphones are spaced 0.4 *m* apart from each other on the floor, and the speaker is placed 0.96 *m* away from the center of the two microphones on the floor of the office. We use spoken digits from the AusTalk database as the sound source. We prepared ten isolated spoken digits (zero-nine) from five speakers, and the spoken digits were played at all the 13 locations. A PC connecting to the two microphones recorded the speech to create the binaural signal database. Additionally, we augmented the database by adding different band-limited noises with different SNRs (between 15 dB and 25 dB), inverting the signals upside down, and stretching the signals in the time domain. Through data augmentation, the database is increased into 11850 samples, where each sample contains an isolated spoken digit from a specific location.

Fig.2 shows the onset detection in the system. The logarithmic speech energy change, $\Delta E\ (t)$, is highly signal and environment related. To detect the signal onset in this experiment, we set the threshold ΔEth to 0.9 and *step* to 125 *ms* in equation (4). For binaural signals, a separated onset time is detected for each cochlea, and the earliest of the two is used as the onset time.

The onset correlogram is generated 90 *ms* after the onset, according to equation (5), i.e., $\Delta t = 90$ *ms*. Fig.3 shows examples of the onset correlograms generated from different azimuthal angles. If there is no delay between the left and right signals, the azimuthal angle is 0°, and there is a strong stripe of

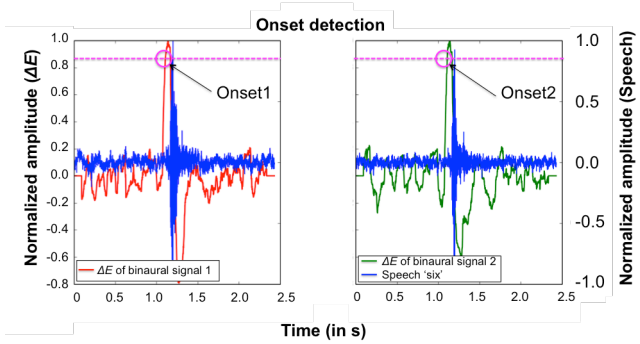

Fig. 2. Onset detection; the log energy change ΔE of the speech (blue wave) is shown in red and green lines, the circles mark the detected onset time for each of the binaural signal. The first onset time t is selected to the onset time.

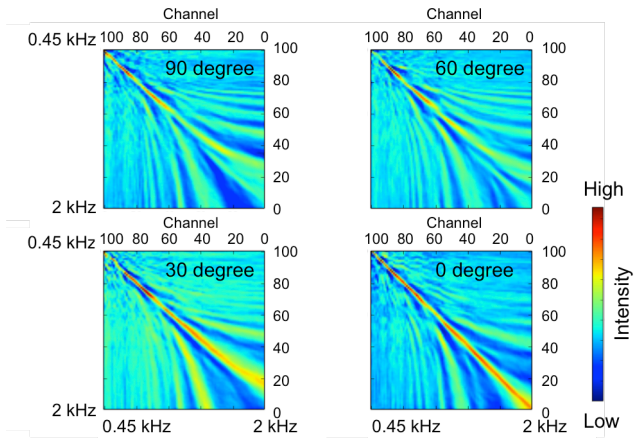

Fig. 3 Correlograms generated from different azimuthal angles of the sound source of speech 'two'.

correlation along the diagonal with symmetric off-diagonal correlation and anticorrelation bands. When there is an ITD between the signals, the correlation stripes are bent towards the cochlea where the signal is delayed, and the off-diagonal bands show an asymmetric structure. At higher frequencies, the correlogram asymmetry is more noticeable.

The 2-D 100×100 zero-mean distributed correlogram is connected to the ELM for the regression analysis. In the simulation, a 2×2 max pooling is used before the ELM so that the size of its input layer is 2500. The hidden layer size is set to 25000, and the *tanh* function is used as the nonlinear activation function. The initial weights have a uniform distribution. The output neuron reports the learned azimuthal angle of the sound source.

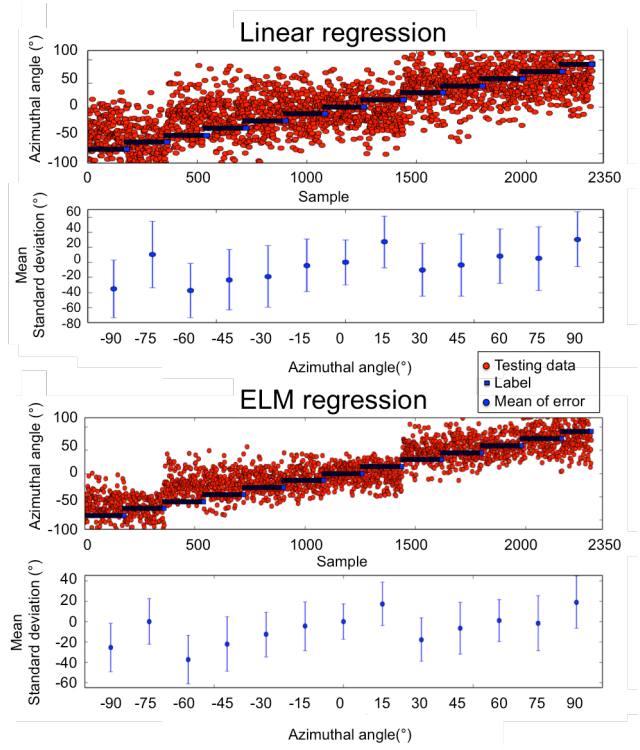

Fig. 4. Linear regression and ELM regression results on the experimental data.

IV. Results

We use the samples from four speakers as the training data (9522) and the samples from the fifth speaker as the testing data (2328). Fig.4 shows the results of testing data distribution, the mean of error and the standard deviation at the 13 locations for the linear and the ELM regression. The ELM shows smaller standard deviation for all locations. The average absolute error is also calculated, and the ELM shows a smaller average absolute error of 12.8° than the linear regression error of 16.5°. Additionally, TABLE I. shows the comparison of this work with prior binaural sound localization systems and algorithms with respect to the approach, setup, input signals, resolution and errors (accuracy). Our system was tested using speech played from positions 15 ° apart in an office environment. This is a more difficult setup than the others in TABLE I, but is closer to a real life indoor scenario. The average absolute error in our tests was larger than in Umbarkar et al., which used wave-counting rather than a trained neural network for localization, but this system only works for localization of pure tones.

TABLE I. Binaural Sound Localization Comparison

	Approach	***Setup***	***Step***	***Error***	***Accuracy***
This work	Correlogram +ELM	Office (Austalk)	15°	12.8°	N/A
Ma et.al [8] (2015)	CCF+ILD+ DNNs+ Head Movements	Room (GRID)	15°	N/A	98.55%
Jin [5] (2014)	ITD+GCC	Seminar room (Speech)	45°	N/A	92.5%
Umbarkar [4] (2011)	Wave count	Room (Sine tones)	15°	1.1° (1 kHz) 3.5° (3 kHz)	N/A
Kugler [19] (2008)	ITD +Spike NN	Room (Non-speech)	30°	N/A	98.7%

V. Conclusions

We have presented a design for a biologically inspired hardware binaural sound localization system for reverberant environments. It exploits the CAR-FAC model to generate 2-D frequency and location dependent correlogram during the sound onset, and the ELM to learn the azimuthal angle of the sound source. The system was evaluated using reverberant input signals and shows an average absolute error of 12.8° in -90° to 90° range. The proposed system is novel in binaural sound localization, and since the correlogram includes both binaural cues and spectral information about the sound source, the design can be extended to other auditory tasks such as sound segregation.

VI. Acknowledgement

The AusTalk corpus was collected as part of the Big ASC project ([20]-[22]), funded by the Australian Research Council (LE100100211). See: https://austalk.edu.au/for details.

REFERENCES

[1] R. F. Lyon, *Human and Machine Hearing -Extracting Meaning from Sound*. Cambridge University Press, 2017.

[2] A. van Schaik and S. Shamma, "A neuromorphic sound localizer for a smart MEMS system," *Analog Integr. Circuits Signal Process.*, vol. 39, no. 3, pp. 267–273, 2004.

[3] V. Y. S. Chan, C. T. Jin, and A. van Schaik, "Adaptive sound localization with a silicon cochlea pair," *Front. Neurosci.*, vol. 4, no. NOV, pp. 1–31, 2010.

[4] A. Umbarkar, V. Subramanian, and A. Doboli, "Low-cost sound-based localization using programmable mixed-signal systems-on-chip," *Microelectronics J.*, vol. 42, no. 2, pp. 382–395, 2011.

[5] J. Jin, S. Jin, S. Lee, H. S. Kim, J. S. Choi, M. Kim, and J. W. Jeon, "Real-time Sound Localization Using Generalized Cross Correlation Based on 0 . 13 μm CMOS Process," *J. Semicond. Technol. Sci.*, vol. 14, no. 2, pp. 175–183, 2014.

[6] T. May, S. van de Par, and A. Kohlrausch, "A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 1, pp. 1–13, 2011.

[7] J. Nix and V. Hohmann, "Sound source localization in real sound fields based on empirical statistics of interaural parameters," *J. Acoust. Soc. Am.*, vol. 119, no. Jan 2006, pp. 463–479, 2006.

[8] N. Ma, G. J. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," in *INTERSPEECH 2015*, 2015, no. September, pp. 160–164.

[9] P. Julian, A. G. Andreou, L. Riddle, S. Shamma, D. H. Goldberg, and G. Cauwenberghs, "A comparative study of sound localization algorithms for energy aware sensor network nodes," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 51, no. 4, pp. 640–648, 2004.

[10] Y. Xu, C. S. Thakur, R. K. Singh, G. Cohen, R. Wang, J. Tapson, and A. van Schaik, "Electronic Cochlea: CAR-FAC Model on FPGA," *IEEE Biomed. Circuits Syst. Conf.*, pp. 1–4, 2016.

[11] R. Wang, G. Cohen, C. S. Thakur, J. Tapson, and A. Van Schaik, "An SRAM-based implementation of a convolutional neural network," in *IEEE Biomedical Circuits and Systems Conference*, 2016, pp. 1–4.

[12] A. Saremi, R. Beutelmann, M. Dietz, G. Ashida, J. Kretzberg, and S. Verhulst, "A comparative study of seven human cochlear filter models," *J. Acoust. Soc. Am.*, vol. 140, pp. 1618–1634, 2016.

[13] D. D. Greenwood, "A cochlear frequency-position function for several species - 29 years later," *The Journal of the Acoustical Society of America*, vol. 87, no. 6. pp. 2592–2605, 1990.

[14] C. A. Mead, X. Arreguit, and J. Lazzaro, "Analog VLSI Model of Binaural Hearing," *IEEE Trans. Neural Networks*, vol. 2, no. 2, pp. 230–236, 1991.

[15] N. Mclachlan and S. Wilson, "The Central Role of Recognition in Auditory Perception : A Neurobiological Model," *Psychol. Rev.*, vol. 117, no. 1, pp. 175–196, 2010.

[16] D. Seidner, "Efficient implementation of log10 lookup table in FPGA," in *IEEE International Conference on Microwaves, Communications, Antennas and Electronic Systems*, 2008, pp. 1–9.

[17] G. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006.

[18] A. van Schaik and J. Tapson, "Online and Adaptive Pseudoinverse Solutions for ELM Weights," *Int. Conf. Extrem. Learn. Mach.*, pp. 1–9, 2013.

[19] M. Kugler, K. Iwasa, V. A. Benso, S. Kuroyanagi, and A. Iwata, "A Complete Hardware Implementation of an Integrated Sound Localization and Classification System Based on Spiking Neural Networks," *Neural Inf. Process.*, vol. 4985, pp. 577–587, 2008.

[20] Burnham, D., E. Ambikairajah, J. Arciuli, M. Bennamoun, C.T. Best, S. Bird, A.B. Butcher, C. Cassidy, G. Chetty, F.M. Cox, A. Cutler, R. Dale, J.R. Epps, J.M. Fletcher, R. Goecke, D.B. Grayden, J.T. Hajek, J.C. Ingram, S. Ishihara, N. Kemp, Y. Kinoshita, T. Kuratate, T.W. Lewis, D.E. Loakes, M. Onslow, D.M. Powers, P. Rose, R. Togneri, D. Tran, and M. Wagner. "A Blueprint for a Comprehensive Australian English Auditory-Visual Speech Corpus". (2009). In The 2008 HCSNet Workshop on Designing the Australian National Corpus, pp.96-107. Sydney: Somerville, MA, USA: Cascadilla Proceedings Project.

[21] Wagner, M., D. Tran, R. Togneri, P. Rose, D. Powers, M. Onslow, D. Loakes, T. Lewis, T. Kuratate, Y. Kinoshita, N. Kemp, S. Ishihara, J. Ingram, J. Hajek, D.B. Grayden, R. Göcke, J. Fletcher, D. Estival, J. Epps, R. Dale, A. Cutler, F. Cox, G. Chetty, S. Cassidy, A. Butcher, D. Burnham, S. Bird, C. Best, M. Bennamoun, J. Arciuli, and E. Ambikairajah. "The Big Australian Speech Corpus (the Big Asc)". (2010). In 13th Australasian International Conference on Speech Science and Technology, edited by M. Tabain, J. Fletcher, D. Grayden, J. Hajek and A. Butcher, pp.166-70. Melbourne: ASSTA, 2010.

[22] Burnham Denis, Dominique Estival, Steven Fazio, Felicity Cox, Robert Dale, Jette Viethen, Steve Cassidy, Julien Epps, Roberto Togneri, Yuko Kinoshita, Roland Göcke, Joanne Arciuli, Marc Onslow, Trent Lewis, Andy Butcher, John Hajek and Michael Wagner. "Building an Audio-Visual Corpus of Australian English: Large Corpus Collection with an Economical Portable and Replicable Black Box". (2011). In Interspeech 2011. Florence, Italy, 2011.