BINAURAL PHONEME RECOGNITION USING THE AUDITORY IMAGE MODEL AND CROSS-CORRELATION

Keith I. Francis[1] *Timothy R. Anderson*[2]

[1]Cedarville College, Cedarville, Ohio, USA
[2]Armstrong Laboratory, Wright-Patterson AFB, Ohio, USA

ABSTRACT

An improved method for phoneme recognition in noise is presented using an auditory image model and cross-correlation in a binaural approach called the binaural auditory image model (BAIM). Current binaural methods are explained as background to BAIM processing. BAIM and a variation of the cocktail-party-processor incorporating the auditory image model are applied in phoneme recognition experiments. The results show BAIM performs as well or better than current methods for most signal-to-noise ratios.

1. INTRODUCTION

This paper presents a binaural phoneme recognizer incorporating the head-related-transfer-functions (HRTFs) of Pösselt et al. [1], the auditory image model of Patterson [2], cross-correlation and a Kohonen [3] neural network. The Pösselt HRTFs are used to produce left and right signals from a source. Each of these signals is corrupted by noise and sent to an auditory image model to produce neural activity patterns for the left and right ear simulation. The neural activity patterns are correlated at a given azimuth and a feature vector is extracted. Such feature vectors are used to train a Kohonen self-organizing feature map which is calibrated as a phoneme recognizer using portions of the TIMIT speech corpus.

A brief background follows succeeded by details of the *binaural auditory image model (BAIM)*; modifications of the *cocktail party processor* (CPP) [4] incorporating the auditory image model (AIM); results from phoneme recognition experiments; and a conclusion.

2. BACKGROUND

2.1. Effectiveness of Binaural Methods

Binaural phoneme recognition using auditory models is more effective than similar monaural methods in noise. For example, DeSimio [5] achieves a 5 dB signal-to-noise ratio (SNR) advantage over monaural processing using a binaural model based on *stereausis*. Bodden and Anderson [4] demonstrate a 20 dB SNR advantage over a similar monaural method using the *Cocktail-Party-Processor*.

2.2. Binaural Fusion by *Stereausis*

DeSimio [5] uses several stages of processing to achieve binaural fusion. First, left and right signals are produced from a source using HRTFs by Wightman and Kistler [6]. Then, the left and right signals are processed through identical cochlear models by Kates [7].

DeSimio [5] uses the *stereausis* technique for binaural fusion. The stereausis processor fuses multi-channel, left and right data streams from the Kates models into a two-dimensional image.

DeSimio extracts feature vectors from the *stereausis* image at selected locations, to improve phoneme recognition in noise. These feature vectors are inputs to a Kohonen self-organizing feature map calibrated as a classifier. DeSimio's system is the first binaural processing system applied to phoneme recognition. *Stereausis* achieves a 5 dB SNR advantage over equivalent monaural processing [8].

2.3. Binaural Fusion by Variants of the Cross-Correlation

Bodden and Anderson [4] use the *Cocktail-Party-Processor*, in an approach similar to DeSimio, to achieve a 20 dB SNR advantage over monaural processing. HRTFs of Pösselt [1] are convolved with speech to produce left and right head-related signals, much like DeSimio. In contrast to the auditory image model of Patterson, used in this work, or the Kates model used by DeSimio, the *Cocktail-Party-Processor* uses a bank of critical-band filters to model the cochlear response. Then, the output from each filter is given to the square root and half-wave rectification functions before submission to the *Cocktail-Party-Processor*.

The *Cocktail-Party-Processor* is based on the inhibited running cross-correlation of Lindemann [9] [10]. Lindemann uses a portion of the contralateral signal to sharpen the initial correlation peak and suppress subsequent correlation peaks. Lindemann suggests the use of the stationary cross-correlation for stationary signals (See Equation 1) and the running cross-correlation (See Equation 2) for non-stationary signals. (In these equations, r stands for data from the right and l stands for data from the left. T is the time constant parameter used to control the impact of the exponential on the correlation.) The *Cocktail-Party-Processor* incorporates the running, inhibited cross-correlation, which is the running correlation computed with modified data from the left*(l)* and right*(r)*. Data modification produces the inhibition. See Lindemann [10] for details.

$$correlation(n,\tau) = \sum_{i=n-N}^{n} r(i+\tau)l(i) \qquad (1)$$

$$correlation(n,\tau) = \sum_{i=-\inf}^{n} r(i+\tau)l(i)e^{-(n-i)/T} \qquad (2)$$

There are two major differences between the stationary and running cross-correlations. The stationary cross-correlation of Equation 1 correlates a limited amount of data with equivalent weight for each product. On the other hand, the running cross-correlation of Equation 2 applies, at least theoretically, to an infinite data set whose products are weighted with an exponential function. Practically speaking, the running cross-correlation applies to a limited data set due to the time constant, T, in the exponential function. Nevertheless, the data is weighted differently than the stationary case if both are applied to the same data.

Finally, the *Cocktail-Party-Processor* produces an azimuth coded vector. The azimuth coded vector consists of the peak values from the inhibited correlation peaks in each critical band. Correlation peaks are produced at a given delay corresponding to the azimuth of the source. Thus, a vector is produced at every time sample interval at a selected azimuth. These vectors are subsequently applied to phoneme recognition with a Kohonen, self-organizing, feature-map (SOFM) calibrated as a phoneme recognizer. Bodden and Anderson's [4] results are plotted in Figure 2 for comparison later in this paper.

2.4. The Auditory Image Model

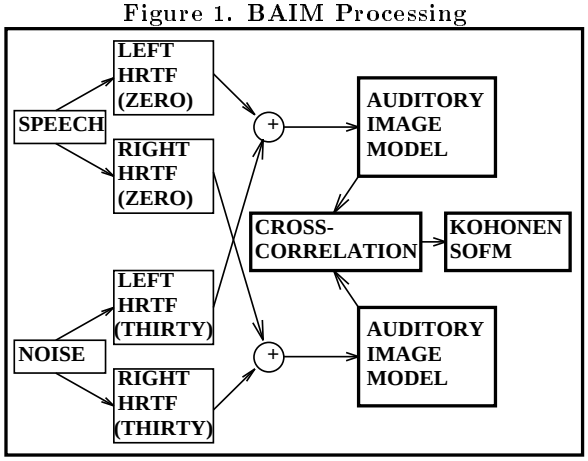
Patterson's *auditory image model* (AIM) [2] [11] simulates the processing of the cochlea and can produce image patterns of sound. His model begins with a bank of gammatone filters with impulse responses comparable to the impulse responses obtained from cats. These filters are better estimates of human filter-shape than the critical bands determined by Zwicker [12]. Specifically, Patterson applies rectification, compression, adaptation, and low-pass filtering after each gammatone filter, to simulate cochlea response, with adaptation in the time domain and suppression in the frequency domain, simultaneously. Consequently, the gammatone filter bank emulates the performance of the basilar membrane and the remaining processes emulate the performance of the inner hair cells and primary auditory nerve fibers of the cochlea.

The *auditory image model* has several options. Functionally, the model produces a response called a neural activity pattern that represents the sum of the activity of the hair cells for the section of the basilar membrane of the cochlea corresponding to the bandwidth of each filter in the gammatone filter bank of the model.

3. BAIM

The *binaural auditory image model* consists of four parts as shown if Figure 1. The first stage uses head related transfer functions of Pösselt et al. [1], to simulate sources at various azimuths relative to the head. The next two stages consist of the auditory image model of Patterson followed by a cross-correlation function. The last stage is a feature extractor. These parts are used for binaural processing in noise in the following way:

Figure 1. BAIM Processing



First, left and right signals are produced. A signal source is convolved with the appropriate HRTFs of Pösselt to produce left and right signals. A speaker at zero degrees is simulated using the HRTFs for zero degrees azimuth. Similarly, a noise source at 30 degrees is produced with the appropriate HRTFs. Thus, each source produces a left and right signal.

Second, noisy speech is produced. This is accomplished by adding all of the left HRTF outputs together and then adding all of the right HRTF outputs together. This produces a noisy left and a noisy right signal.

Third, the noisy signals are sent to auditory image models. The auditory image model produces neural activity patterns from each gammatone filter in the filter bank. (See [2] for model details). Thus, the auditory image model simulates the performance of the cochlea in the inner ear. In this manner, a data stream is produced for a selected number of frequency bands in the range of human hearing. In this work, 18 channels are used for each ear.

Fourth, the outputs from the auditory image models are correlated. These correlations are done for each channel across corresponding channels of the left and right neural activity patterns generated by the auditory image models. The correlation is done over a window which is 33 points wide. Thus, a stationary cross-correlation, shown in Equation 1, is used instead of a running, inhibited cross-correlation. The correlation window is advanced through the left and right signal, one sample point at a time, in each channel; consequently, a correlation result is produced for every sample point in every channel for the length of the utterance.

After the correlation is performed, a feature vector is extracted across the channels for a given azimuth using the correlation results from each channel. Since a stationary correlation is performed in each channel across right and left neural activity patterns, the maximum correlation is expected at $\tau = 0$ for a speaker at zero degrees azimuth relative to a person's head. (See Equation 1.) Therefore, for these experiments, the correlations of the left and right neural activity patterns are calculated at $\tau = 0$ to produce the feature vectors for a speaker at zero degrees to be sent to the next stage of processing. Please note for future reference: this is called the *no head adaption* case. It is possible

Figure 2. Cocktail-Party-Processor vs BAIM

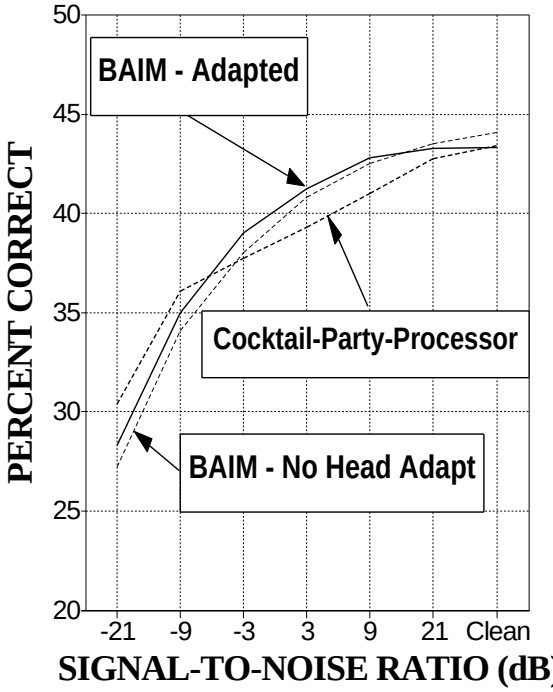
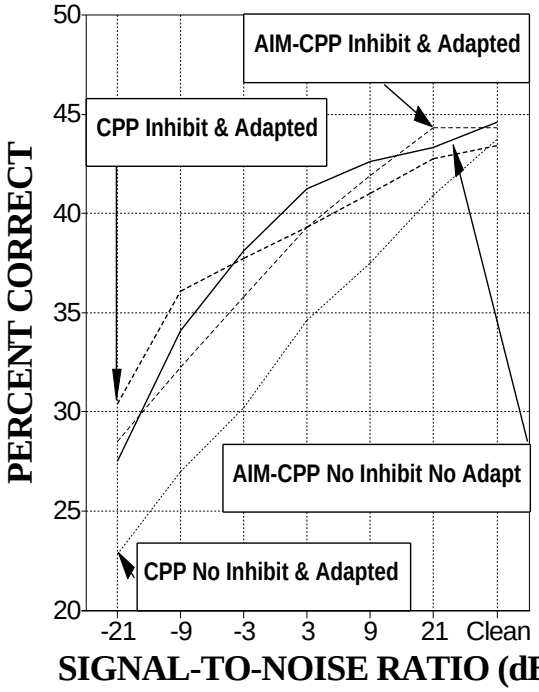


Figure 3. Cocktail-Party-Processor Modifications



to gain an additional improvement in phoneme recognition by choosing a more precise τ slightly different due to the impact of head shape in the HRTFs. This is done by determining the τ of maximum correlation due to an impulsive sound at zero degrees for each frequency band. Then, these τ's are used in subsequent correlation calculations for feature vectors. This later case is referred to as the *head adapted* case in the discussion that follows.

Finally, the feature vectors are used with a Kohonen self-organizing feature map calibrated as a classifier. This classifier is trained on the feature vectors described above using the *leave-one-out-method* described by DeSimio [5] [8] and Bodden and Anderson [4] for ten utterances, each from ten speakers, selected from the TIMIT speech corpus. Then the trained classifier is used in the tests of the system.

Using the method above, tests were performed for seven signal-to-noise ratio cases: -21 dB, -9 dB, -3 dB, 3 dB, 9 dB, 21 dB and clean (no noise). The results are shown in Table 1, Table 2 and Figure 2 for the *head adapted* and *no head adaption* cases. Notice the slight improvement in the *head adapted* case. These results are also contrasted with Bodden and Anderson's *cocktail-party-processor* [4] in Figure 2.

Table 1. BAIM: No Head Adaptation
% Correct vs SNR

-21dB	-9dB	-3dB	3dB	9dB	21dB	Clean
27.2	34.1	38.0	40.8	42.5	43.5	44.1

4. VARIATIONS OF THE COCKTAIL-PARTY-PROCESSOR

4.1. The Uninhibited Cocktail-Party-Processor

An *Uninhibited Cocktail-Party-Processor* can be produced by eliminating inhibition in the running cross-correlation of the *Cocktail-Party-Processor*. The *Cocktail-Party-Processor* is based on the running, inhibited cross-correlation. This has been discussed above in Section 2.3. Nevertheless, the inhibition may be eliminated to produce an *Uninhibited-Cocktail-Party-Processor*. The result is a binaural processor based on the simple auditory model used by Bodden and Anderson coupled with the running cross-correlation for binaural fusion.

The *Uninhibited Cocktail-Party-Processor* was tested in the same noise conditions as above with head adaptation applied. The results are shown in Table 3. The results are contrasted with Bodden and Anderson's *cocktail-party-processor* [4] and the *AIM-Cocktail-Party-Processor* in Figure 3. The *AIM-Cocktail-Party-Processor* is discussed below.

Table 2. BAIM: Head Adapted
% Correct vs SNR

-21dB	-9dB	-3dB	3dB	9dB	21dB	Clean
28.3	35.0	39.0	41.2	42.8	43.2	43.3

Table 3. CPP No Inhibit & Adapted
% Correct vs SNR

-21dB	-9dB	-3dB	3dB	9dB	21dB	Clean
22.8	27.0	30.2	34.6	37.5	40.9	43.7

4.2. AIM and the Cocktail-Party-Processor

The auditory image model and the binaural processor used in the *cocktail-party-processor* can be combined to produce a *AIM-Cocktail-Party-Processor* (AIM-CPP). To visualize this combination, replace the correlation process in BAIM, see Figure 1, with the binaural processor used in the *Cocktail-Party-Processor*.

The *AIM-CPP* was tested with two parameter sets in the same noise conditions as before. The results are contrasted with Bodden and Anderson's *cocktail-party-processor* [4] in Figure 3. In the first series of tests, inhibition and head adaptation were disabled. Head adaptation was not used in this case because it was not clear how to determine the head adaptation for this type of configuration. (The numerical results are shown in Table 4.) Next, the *AIM-CPP* was tested using the exact parameters used by Bodden and Anderson's *cocktail-party-processor* including the head adaption normally used with the *cocktail-party-processor*. This is called the *Inhibited & Adapted* case. (The numerical results are shown in Table 5.)

Table 4. AIM-CPP No Inhibit No Adaptation

% Correct vs SNR

-21dB	-9dB	-3dB	3dB	9dB	21dB	Clean
27.5	34.1	38.1	41.2	42.6	43.3	44.6

Table 5. AIM-CPP Inhibited & Adapted

% Correct vs SNR

-21dB	-9dB	-3dB	3dB	9dB	21dB	Clean
28.5	32.2	35.8	39.3	41.9	44.3	44.3

5. DISCUSSION AND CONCLUSIONS

BAIM meets or exceeds the performance of previous binaural phoneme recognition systems for most signal-to-noise ratios. Considering the seven noise conditions, BAIM performed as well as CPP for -9 dB, -3 dB, 3 dB, 21dB and clean cases based on the two sided *t-test* with a 95% confidence level with nine degrees of freedom. In the 9 dB case BAIM was significantly better. In the -21dB case CPP was significantly better.

AIM-CPP, uninhibited and unadapted, meets or exceeds the performance of previous binaural phoneme recognition systems in four of the seven noise conditions without inhibition and head adaptation. CPP is significantly better for the worst noise conditions: -21dB and -9 dB. AIM-CPP, uninhibited and unadapted, is significantly better at 9 dB. In the remaining cases, there is no significant difference using the same *t-test* as above.

The benefits of the running correlation are not apparent. BAIM, without head adaptation, is not significantly different from AIM-CPP without inhibition and head adaptation. Since the running cross-correlation has many possible weightings depending on the parameter T (See Equation 2) a conclusive statement cannot be made based on the data presented; nevertheless, the benefits of the running cross-correlation are suspect.

Thus, BAIM, which combines the auditory image model and stationary cross-correlation, produces a phoneme recognition rate equal or better than the *Cocktail-Party-Processor* for most signal-to-noise ratios. Also, the auditory image model combined with a running cross-correlation without inhibition performs nearly as well as the *Cocktail-Party-Processor*. This demonstrates the benefits of binaural processing in noise and the potential for further advances in this area with sophisticated auditory models. Future work will incorporate the auditory image model with a new fusion technique based on coincidence detection of neural activity patterns.

REFERENCES

[1] C. Pösselt, J. Schröter, M. Opitx, P. Divenji, and J. Blauert, "Generation of binaural signals for research and home entertainment," in *Proc. 12th Int. Congr. on Acoustics*, 1986.

[2] R. D. Patterson and J. Holdsworth, "A functional model of neural activity patterns and auditory images," in *Advances in Speech, Hearing and Language Processing*, Reading, Massachusetts: JAI Press, 1991.

[3] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, Sept. 1990.

[4] M. Bodden and T. R. Anderson, "Cocktail party speech recognition," in *EuroSpeech95*, (Madrid, Spain), Sept. 1995.

[5] M. P. DeSimio, *Speaker independent phoneme recognition with a binaural auditory model*. PhD thesis, University of Dayton, Dayton, Ohio, 1993.

[6] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listeng. i: Stimulus synthesis," *J. Acoust. Soc. Am.*, vol. 85, no. 2, pp. 858–867, 1989.

[7] J. M. Kates, "A time-domain digital cochlear model," *IEEE Trans. on Signal Processing*, vol. 39, no. 12, pp. 2573–2592, 1991.

[8] M. P. DeSimio, T. R. Anderson, and J. J. Westerkamp, "Phoneme recognition with a model of binaural hearing," *IEEE Transaction on Speech and Audio Processing*, vol. 4, no. 3, pp. 157–166, 1996.

[9] M. Bodden, "Modeling human sound-source localization and the cocktail-party-effect," *acta acustica*, vol. 1, pp. 43–55, February/April 1993.

[10] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. i. simulation of lateralization for stationary signals," *J. Acoust. Soc. Am.*, vol. 80, pp. 1608–1622, Dec. 1986.

[11] R. D. Patterson and M. H. Allerhand, "Time-domain modeling of peripheral auditory processing: A moduar architecture and a software platform," *J. Acoust. Soc. Am.*, vol. 98, 1995.

[12] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *J. Acoust. Soc. Am.*, vol. 33, p. 248, 1961.