THE COCKTAIL PARTY LISTENER

A. R. Kian Abolfazlian & Brian L. Karlsen

Computer Science Department
Aarhus University

ABSTRACT - A complex computational model of the human ability to listen to certain signals in preference of others, also called the cocktail party phenomenon, is built on the basis of surveys into the relevant psychological, DSP, and neural network literature. This model is basically binaural and as such it makes use of both spectral data and spatial data in determining which speaker to listen to.

The model uses two neural networks for filtering and speaker identification. Results from some experimentation with type and architure of these networks is presented along with the results of the model.

These results indicate that the model has a distinctive ability to focus on a particular speaker of choice.

INTRODUCTION

This paper deals with a very interesting psychological phenomenon: The cocktail party phenomenon. In short we have attempted to construct a computer model capable of concentrating on the speech of specified speakers when several people are speaking simultaneously. Readers, who are interested in the details of our model or in the extensive literature surveys produced in the preliminary stages of this project, are referred to (Abolfazlian & Karlsen 1994, or Karlsen & Abolfazlian, 1994).

One of the most prominent theories of such auditory processing is the one promoted by Bregman (1990). Basically, his idea is that the objects of attention are the auditory events, called *streams*, which are sounds that are perceived as being a whole in some sense, e.g. a singer and an accompanying piano. The sounds that we receive are processed in two different stages: the primitive process, and the schema-based process. The primitive process makes use of cues such as pitch, formants, sound source localization, intensity and others to form units and to group these units together in both the dimensions of time and spectrum. This is what Bregman calls sequential integration and simultaneous integration respectively. Once these groupings have been made one needs to decide which group deserves to get attention. This task is handled by the schema-based process, which on the basis of a mental organization of information is capable of selecting evidence out of a mixture, which has not been subdivided by the primitive process. So the primitive process basically partitions the evidence while the schema-based process selects from the resulting partitions.

Digital signal processing (DSP) is usually a must when handling sounds computationally. In DSP one can basically make use of two types of method for analysis of signals: the nonparametric and the parametric methods. The nonparametric methods include Fourier analysis, and these methods are very good at deriving the objective measures inherent in a signal as they generally ignore the sound producing mechanism, and they do therefore not assume anything about the acoustical-mechanical structure of the signal. The parametric methods on the other hand allow compact representations, generally by means of a reduced number of coefficients, as in Linear Predictive Coding, and take advantage of certain *a priori* information; for example, a known production model of the sound (e.g. the voice) has a parametric representation. Parametric methods are extremely interesting to computer processing, as they allow an in depth analysis of numerous factors about the sound ranging from the psychoacoustics to the physical structure of the sound generator. Consequently, parametric methods are very useful tools from an analysis point of view.

Artificial neural networks (ANN) are usually very good at classifying a mixed set of inputs, so it is only natural to try to apply these computational architectures to signal processing as well. From our point of view there are two overall types of neural networks; those that can handle temporal context and those that cannot. The latter of these is usually referred to as feedforward networks. These networks are capable of subdividing any set of static patterns into classes, if the network is built and taught properly. The former type of neural network is a mixed set consisting of time-delay neural networks and sequential neural networks (SNN) (Hertz et al., 1990). The SNN have limited feedback connections in the architecture, limited in the sense that only certain units have these connections and the weights on these connections may be fixed. One example of such a network with short-term memory context units is the one put forth by Jordan (1986).

In the following sections of this article we shall present our model, the experiments that we performed, and the results that we obtained. Finally, we will give some pointers to future questions that need to be investigated.

THE STIMULI

The stimuli were recorded at a sampling rate of 44.1 kHz using a dummy head (Brüel & Kjær Head and Torso Simulator WH2511), a microphone power supply (Brüel & Kjær Two Channel Power Supply), and a DAT recorder (Studer D780). The recordings were made with 5 different Danish speakers placed at 3 different positions in two different rooms yielding a total of 2 sets of 15 recordings. Once these recordings had been made they were transferred onto a NeXT for editing and DSP. Here the soundfiles were first cut down to a uniform size in each set. Then the soundfiles were mixed so as to obtain all physically possible combinations of speakers in different positions in the two sets (a total of 120 combinations in each set). These mixes were then used as the raw data of the model. However we had to downsample the soundfiles to 22.05 kHz in order to be able to use the DSP package called CARL (Loy, 1994).

THE MODEL

The model consists of a number of signal processing tools and an ANN for identifying speakers from their fundamental frequencies and formants. The model is only concerned with specific speakers and not with words. The clues that it makes use of are the aforementioned frequencies and the speakers azimuth angle relative to the listener. Basically, the model consists of three major units: the spectral pathway, the spatial pathway, and the selective attention unit.

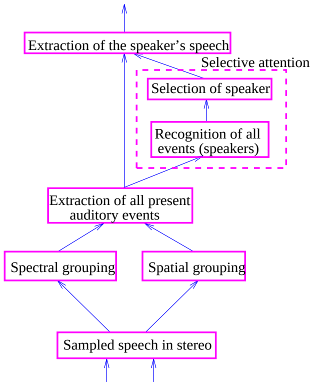
Extraction of the speaker's speech
Selective attention
Selection of speaker
Recognition of all events (speakers)
Extraction of all present auditory events
Spectral grouping
Spatial grouping
Sampled speech in stereo

Figure 1: The coctail party listener model

The Spectral Pathway

In this unit the raw speech signals from both channels are first transformed into Linear Predictive Coding (Makhoul, 1975). The resulting coefficients are then used to derive the fundamental frequency as well as the formants. The LPC is carried out using an all-pole model and the covariance method (Markel & Gray, 1976), as we are here dealing with speech which can be considered to be a quasi-stationary type of signal. The fundamental frequency of each speaker is computed using the autocorrelation of the LPC models error signal (Saito & Nakata, 1985), and the formants are computed by solving the equation

$$A(z) = z^p + \alpha_1 z^{p-1} + \alpha_2 z^{p-2} + \cdots + \alpha_{p-1} z + \alpha_p = 0 \qquad (1)$$

where the roots correspond to the poles of the all-pole LPC model with transfer function $H(z) = 1/A(z)$. For each time-slice these data are then fed to a sequential neural network (Jordan, 1986) which has been trained to disect separate auditory events (e.g. speakers). On the basis of the network output it is then possible to estimate how many auditory events are present and what their separate signals are. In short the network is actually performing an analysis which integrates spectral components both simultaneously and sequentially as Bregman has proposed (1990).

The Spatial Pathway

This part of the model was heavily inspired by Blauert's ideas of "Spatial Hearing" (1983). Here the raw speech signals are first bandpass filtered using a bank of critical bandpass filters. Thereafter a lowpass filter is applied to all the band signals from both channels in order to derive the envelope of the signals. For each pair (left and right) of these lowpassed band signals a set of specialized cross-correlation coefficients are then computed as follows:

$$\Phi_i = \sum_{n=0}^{N-i-1} x_n y_{n+i} G(N-i-1-n) \mathsf{sign}(x_n - y_{n+i}) L(x_n - y_{n+i}), \qquad (2)$$

where x is the left signal, y is the right signal, N is the number of samples over which to average, $i = 0, \ldots, T-1$ where $T = 13$ samples ($\simeq 0.6$ ms) is the number of delays, and

$$G(s) = \begin{cases} e^{-s/\tau_{RC}} & \text{for } s \geq 0 \\ 0 & \text{for } s < 0 \end{cases}$$

is a weighting function that gives less and less significance the farther the values lie in the past ($\tau_{RC} = 28$ samples $\simeq 1.25$ ms) and

$$L(s) = -e^{-s^2/M} + 1.1 \qquad (3)$$

is the level difference weighting function we have chosen. This function is maximal when the absolute value of s is maximal, otherwise it drops off monotonically as s goes to zero as required by Blauert. M is chosen so that it fits the order of magnitude of s, in this case 5.0×10^{-10}. On the basis of these coefficients the position on the axis of delay where the maximum correlation between the left and the right channel in each band is found. This delay time corresponds to the estimated azimuth angle of the sound source relative to the listener. This means that each band gets one "vote" in estimating the angular positions of all the present sound sources, but we have found it necessary to introduce a threshold value when counting the "votes": If more than 10% of the total number of bands point to the same value on the time delay axis, we say that there is a genuine sound source, otherwise these votes are regarded as spurious and are ignored when estimating the number of speakers. By averaging the estimates over several time frames we obtain a greater reliability.

The Selective Attention Unit

The previous two analyses combine to yield a determination of what auditory events are present in the input currently. In principle the results of the two pathways can be in conflict with each other. Such a conflict is then resolved by preferring the estimates of the spectral pathway as the psychological evidence shows that we humans do not assign such a big importance to the results of the spatial pathway, simply because it is more uncertain due to possible interference from all kinds of physical circumstances (e.g. echo, obstructing objects, noise etc.). The auditory events determined by the output of the filtering ANN in the spectral pathway is then used one at a time as input for the identification ANN in an attempt to recognize the speakers. If a speaker is identified by the ANN the next step is to determine whether this particular speaker is of interest to the model or not. This is simply done by checking a priority queue to see if the speaker is the one with top priority at the moment. Once all this is done the last remaining step takes care of outputting only the fundamental and the formants of the interesting speaker.

SPEAKER IDENTIFICATION USING ANN'S

Two basic architectures are compared in use as artificial neural networks for identifying the speakers on the basis of the different frequencies. These two architectures are a basic feedforward network and a recurrent network with context units in Jordan style (Jordan, 1986). From a theoretical standpoint the recurrent network should in principle be better at solving the task of speaker identification than the feedforward network, simply because the recurrent network has a kind of short-term memory incorporated in the architecture. Therefore the recurrent network has a broader base on which to make a classification, in our case extending 200 ms back in time. The feedforward network on the other hand must base a classification solely on one time-frame, in our case corresponding to 15 ms.

The actual networks that we used for our experiments were a 24-12-6 feedforward network trained by using standard backpropagation (Rumelhart et al., 1986) and off-line learning, and a 24-12-6-6 sequential network also trained by using backpropagation and off-line learning. The parameters used in both cases were a learning rate $\eta = 0.25$, a momentum term $\alpha = 0.9$, and a maximum number of epochs of 1501. Also in both cases the network was trained from scratch 10 times, each time with a new random intialization of the weights of the network between -1 and 1. All the unit activations were initialized to 0 except from the context units in the recurrent network which were initially set to 0.5. During training the target values were copied back into the context units of the recurrent network and not the actual output values of the network.

The inputs of the networks were normalized formants and fundamentals of the individual present speakers, and the targets were localized binary codings, in which one speaker would turn on exactly one unit in the output layers of the networks.

RESULTS OF THE MODEL

Spatial Pathway

Running the 120 different mixes through the spatial pathway produced some quite interesting results. In figure 2 and figure 3 one can see the average estimated number of speakers over time for each mix as the completed line while the dashed line marks the actual number of speakers.
We did not test the model on one person speaking, because this was considered to easy for the model to esimate. If you threshold the output values of the spatial pathway at 1.5 and 2.5 it is possible to see that the model makes correct estimates in all the situations.
Also in figure 4 and in figure 5 we have depicted the root mean square error of the estimates that the model makes on the different mixes. These graphs basically show a quite small error in relation to the correct number of speakers. The average error made by the model is 13.2% and 14.4% in room 1 and room 2 respectively. If you isolate the error in the two situations when 2 and 3 people are speaking, you get an average error for 2 speakers in room 1 of 16.4% and 9.9% for 3 speakers in room 1. In room 2 the corresponding ratios are 17.3% for 2 speakers and 11.5% for 3 speakers. From this it is clear that the model is not equally good at estimating 2 and 3 speakers, however due to time-constraints we have not been able to test if this difference is statistically significant. But we have computed the correlation coefficient between number of speakers and root mean square error. This gave -0.549 for room 1 and -0.506 for room 2. It is clear from this that the root mean square error grows as the number of speakers decrease, but what is the cause of this? Does it depend on the number of male or female speakers? This was also tested by computing the correlation coefficients between the number of male/female speakers in each mix and the root mean square error. In room 1 this resulted in -0.247 for male speakers and -0.172 for female speakers. In room 2 the results were -0.337 for male speakers and -0.043 for female speakers. As can be seen, all these correlation coefficients are negative indicating that it is only the number of speakers that is responsible for the reverse relationship to root mean square error and not the sex of the speakers.

Spectral Pathway

Training the filtering network on the set of data obtained from our recordings in room 1 as described above, we found that the sequential network needed about 4500 epochs to be able to perform the task in a satisfactory manner. In figure 6 the networks performance, averaged over the 10 runs, on the training pattern can be seen as the top curve. This curve shows a quite slow, but reasonable, climb in performance as the number of epochs grow, ending up at about 93.1% performance. As the bottom curve in figure 6 the corresponding performance on the test pattern set from room 2 can be found. We have, as the reader can gather from this figure, tested the entire test set after every learning epoch making it possible for us to see the development in the networks ability to generalize as training progresses. The networks final performance on the test set was 89.7%.
It should be clear that the networks ability to generalize to the test patterns follows the rise in performance for the training patterns quite closely as training progresses. Considering the non-static nature of the inputs and targets this is quite impressive. Another thing which is impressive is how well the sketched method for determining the number of speakers on the basis of the network output works. This can for example be seen in figure 7. Basically, the network determines the number speakers absolutely correct in all cases in the training room (room 1) and only makes few mistakes in the test room (room 2).

Selective Attention

As the reader may remember we had decided to compare two types of ANN architectures for the identification part of the selective attention unit. In figure 8 the performance of the feedforward network averaged over 10 runs on both the training set and the test set can be seen as the top and bottom curve respectively. The two curves clearly follow each other rather nicely. After 1500 epochs where we terminated training, the performance on the training set was 93.1% and correspondingly 92.3% on the test set. This indicates a very good ability to generalize to other examples of speech from each speaker, which again seems to point towards the conclusion that the network has learned some kind of significant speech characteristic of each speaker.
Like the performance curves of the feedforward network followed each other rather closely, so does the corresponding curves of the sequential network. But there is one big difference: the performance curves of the sequential network rise slower than the ones for the feedforward network as can be seen in figure 9. However the final performance of the sequential network after 1500 epochs is 97.4% and 94.7% for the training set and test set respectively. This means that the sequential network is actually significantly better than the feedforward network at the identification task after being trained in exactly the same way on exactly the same patterns.We chose to keep our priority queue static.

CONCLUSION

All in all one can conclude from the above results that our model is quite apt at performing its task.
Our model seems to predict a better performance on determining number of speakers when there are 3 speakers present than when there are 2. There do not to our knowledge exist any psychological experiments which have tested this.
Futhermore, we have determined that a sequential network is more apt at identifying speakers than a feedforward network is. So we recommend that researchers working with speaker identification use some kind of neural network with feedback connections implementing a short-term memory. This is clearly superior to no memory feedforward networks.
Finally, one should keep in mind that the way the task of the model is performed might not be possible to state explicitly in terms of rules or words but only in terms of examples. This is a statement which applies generally to the area of auditory modelling. In our opinion researchers in this field are in general far to focused on determining the acoustical features that governs this and that piece of the auditory system and not so interested in the actual functionality of the system.
For simplicity's sake we have chosen to ignore all other possible dimensions of the input than fundamental frequency and formants, and in line with the above argument it should in fact be avoided to split the input up into a lot of more or less artificial features. It would be much more functionally satisfying if one could simulate the processes of the auditory system by using some kind of self-organizing system on the raw data. However, these kinds of systems are not quite so advanced yet that this is a possible option, and one is therefore forced to handle all the preprocessing explicitly which necessarily leads to measuring certain dimensions.
We believe that the overall functionality of the model is close to being cognitively correct as is Bregman's theory on auditory scene analysis, so in principle it is only a matter of finding the right way to implement it. This includes finding new ways of processing sound, and as argued above this is one of the primary problems of cognitive auditory models of today. Once this problem is solved we will probably be able to built specialized hearing aids for the elder that will do away with their cocktail party problem, but in our opinion the technique for this is not completely present yet.

REFERENCES

Abolfazlian, A. R. K., & Karlsen B. L. (1994) 'The Cocktail Party Listener', Technical report DAIMI PB 482, Computer Science Department, Aarhus University, ISSN 0105-8517

Blauert, J. (1983) 'Spatial Hearing - The Psychophysics of Human Sound Localization', MIT Press

Bregman, A. S. (1990) 'Auditory Scene Analysis - The Perceptual Organization of Sound', MIT Press

Hertz, J., Krogh, A., & Palmer, R. G. (1990) 'Introduction to the Theory of Neural Computation', Addison-Wesley

Jordan, M. I. (1986) 'Attractor Dynamics and Parallelism in a Connectionist Sequential Machine', in *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, pp. 531-546, Erlbaum

Karlsen, B. L., & Abolfazlian, A. R. K. (1994) 'On Selective Attention in the Auditory Domain: A Hybrid Cocktail Party Listener', Master's Thesis, Computer Science Department, Aarhus University, ISSN 0106-9969

Loy, G. (1994): 'Introduction to CARL Programming', Computer Audio Research Laboratory, Center for Music Experiment, University of California at San Diego

Makhoul, J. (1975) 'Linear Prediction: A Tutorial Review', *Proc. IEEE*, vol. 63, pp. 561-580

Markel, J. D., & Gray, A. H. Jr. (1976) 'Linear Prediction of Speech', Springer-Verlag

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986) 'Learning Internal Representations by Error Propagation', in *Parallel Distributed Processing*, Rumelhart, D. E., McClelland, J. L. and The PDP Research Group, MIT Press

Saito, S., & Nakata, K. (1985) 'Fundamentals of Speech Signal Processing', Academic Press

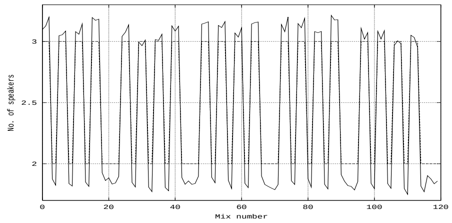

Figure 2: Estimates on number of speakers in room 1

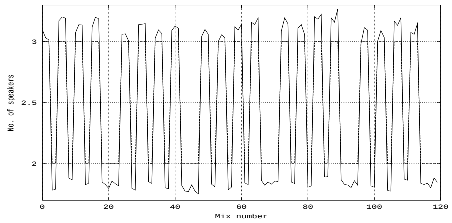

Figure 3: Estimates on number of speakers in room 2

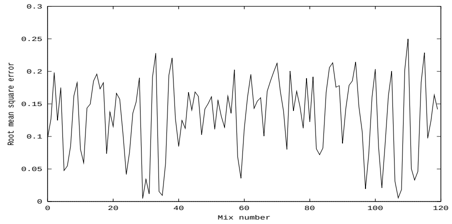

Figure 4: Root mean square error on the estimates in room 1

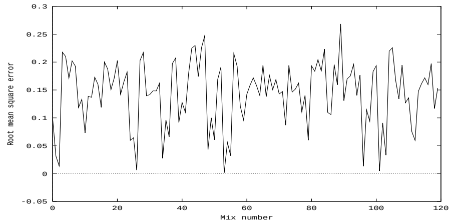

Figure 5: Root mean square error on the estimates in room 2

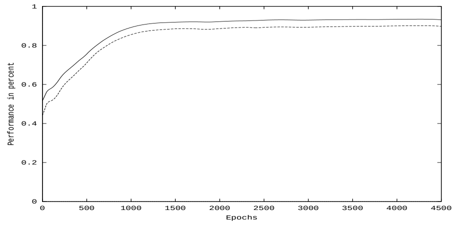

Figure 6: Performance of the filtering ANN during training (the top curve) and testing (the bottom curve)

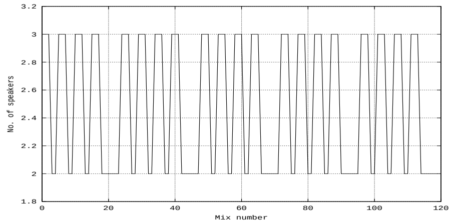

Figure 7: Number of speakers in room 1 estimated by the filtering network

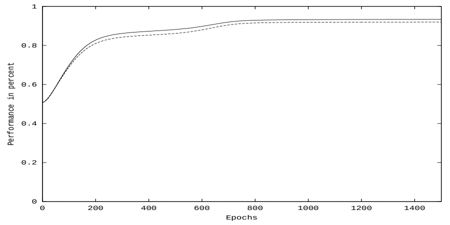

Figure 8: Performance of the feedforward identification ANN during training (the top curve) and testing (the bottom curve)

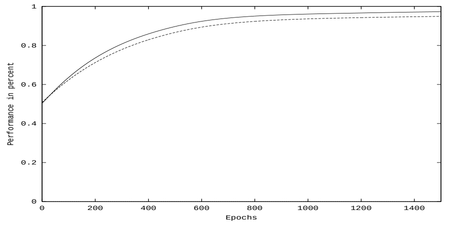

Figure 9: Performance of the sequential identification ANN during training (the top curve) and testing (the bottom curve)