The Role of Top-Down Attention in the Cocktail Party: Revisiting Cherry's Experiment after Sixty Years

Letizia Marchegiani*†, Seliz G. Karadoğan*, Tobias Andersen*, Jan Larsen* and Lars Kai Hansen*‡
*DTU Informatics, Technical University of Denmark, DK-2800, Kgs. Lyngby, Denmark
Email: malet,seka,ta,jl,lkh@imm.dtu.dk
†Department of Computer and System Sciences,
Sapienza, University of Rome, 00185 Roma, Italy
‡Department of Signals and Communications,
University Carlos III, Madrid, Spain

***Abstract*—We investigate the role of top-down task drive attention in the cocktail party problem. In a recently proposed computational model of top-down attention it is possible to simulate the cocktail party problem and make predictions about sensitivity to confounders under different levels of attention. Based on such simulations we expect that under strong top-down attention pattern recognition is improved as the model can compensate for noise and confounders. We next investigate the role of temporal and spectral overlaps and speech intelligibility in humans, and how the presence of a task influences their relation. For this purpose, we perform behavioral experiments inspired by Cherry's classic experiments carried out almost sixty years ago. We make participants listen to a mono signal consisting of two different narratives pronounced by a speech synthesizer under two different conditions. In the first case, participants listen with no specific task, while in the second one they are asked to follow one of the stories. Participants report the words they heard by choosing from a list which also includes terms not present in any of the narratives. We define temporal and spectral overlaps using the ideal binary mask (IBMs) as a gauge. We analyze the correlation between overlaps and the amount of reported words. We observe a significant negative correlation when there is no task, while no correlation is detected when a task is involved. Hence, results that are well aligned with the simulation results in our computational top-down attention model.**

I. Introduction

The cocktail party is an often used analogy in machine learning and signal processing, referring to the situation in which multiple signals are mixed and the aim is to separate these, or to recover at least one of the signals in the mixture. In the case of audio mixtures humans are very efficient cocktail party solvers, using a multitude of cues including spatial, spectral, and content as was demonstrated in the famous experiments carried out by Colin Cherry almost sixty years ago [1].

We note that many machine learning applications face 'cocktail parties'. In biomedical signals noise and confounders are often structured and share features with the wanted signal, hence, prior information about the signals and the ways they are mixed plays an important role [2]. Modern telecommunication systems are critically dependent on the ability to recover individual signals from spread spectrum mixtures [3]. Most of the general methods use low level statistical properties of the signals, such as independence [4], or other simple distributional assumptions.

We are interested in the audio cocktail party problem, in particular, the 'hard version' considered in [1], in which, conventional audio cues (spatial and spectral) are removed and the solution is based on high level features related to content. As discussed in [1], this problem can help shed light on the mechanisms applied by human cognition. We are particularly interested in the influence of top-down attention [5], [6].

The audio cocktail party problem is also of great practical importance, e.g., in the transcription of multi-speaker conferences, meetings, seminars and in dialogue systems of robots operating in complex acoustic scenes. Automatic speech recognition procedures need to isolate the voice of interest within confounding sounds and voices around and to track it, to be able to recognize the words pronounced ([7], [8]).

Experiments have been conducted to investigate which characteristics of the auditory scene could help the segregation process of a mixture of stimuli and in which way they can influence each other and the human ability to discriminate within the different signals. The majority of these experiments make use of pure tones, see e.g. [9]). But there are also cases in which human auditory behaviours are tested in situations with multiple speakers, see e.g., [9], [10] and [11]. Sound fission seems to be more pronounced with the voices of the same gender [11], with sounds emitted from close positions in the scene or sounds having enough difference in their fundamental frequencies, in their phase spectrum or in their intensities [10]. Meanwhile, also the vocal tract size, accent or other prosodic features can change the complexity of the grouping of signals belonging to the same stream [9]. For a more complete review, see [12] and the more recent [13].

Bregman and others ([9], [14], [15]), argue that the way in which stimuli are perceived as part of the same flow (coming from the same acoustic source), and the proficiency in selecting just one of these flows and understanding its content, is widely affected by attention, both on a bottom-up and a top-down perspective. It should also be considered that, in fact, the segregation ability is a learned skill and is improved by experience. In psychoacoustics masking refers to the effect that one signal prohibits the other from being detected. Brungart et al. make a distinction between energetic and informational masking: *"Traditional energetic masking occurs when both utterances contain energy in the same critical bands at the same time and portions of one or both of the speech signals are rendered inaudible at the periphery. Higher-level informational masking occurs when the signal and masker are both audible but the listener is unable to disentangle the elements of the target signal from a similar-sounding distracter"* [16].

In order to better understand top-down attention, and how it may modulate informational masking effects we return to Cherry's basic experimental setup, i.e., a listening experiment in which we investigate participants' ability to hear individual naturally sounding speech signals in a mixture with reduced spatial and speaker cues. In particular, we present the audio signal to the listener as a *monaural mixture* of two different narratives uttered by a *speech synthesizer*

(TTS project at AT&T Labs Research [17]), using the same virtual speaker. In this way, we eliminate cues to separation related to different spatial locations of the sound sources, different accents, different genders of the speakers etc. However, we still want the speech to sound natural, hence, the speech synthesizer does produce prosodic speech which also provides a cue to stream separation and tracking [18]. It is known that the introduction of this kind of voice modifications effects detectability [19], however, we expect these effects to be reduced compared to conventional human speech. To further reduce energetic masking, we equalized the total energy of the mixed signals.

In order to reduce basic semantic masking effects we opted for narratives with little expected interest to the listeners. In particular we chose as excerpts from neutral texts used in preparation for the TOEFL (Test of English as a Foreign Language) test [20]. These texts are more coherent and naturally sounding than the short command sentences used in [16].

We have recently proposed a computational mechanism for task driven top-down attention based on a generative statistical model of inputs, corresponding to a 'gist' of the scene and to potential elements for attention, and task labels corresponding to possible actions chosen based on gist and attended inputs ([5], [6]). The notion of gist refers to unspecific inputs generated in part by bottom-up attention [21]. This model can be used to make predictions about the influence of confounders on task labeling performance, in both strongly task dependent attention and under weakened task influence. Thus we have designed experiments so that they test both of these cases. In the first set of experiments (we call it *undirected attention* (UNDIR)), the subjects do not have any specific task rather than simply aim to hear as many words as possible, thus they may follow any of the two narratives, while, in the second set of experiments (*directed attention* (DIR)), they are asked to focus on just one of the two narratives (the choice about which of them is left to the subject).

To test the relative influence of top-down and bottom-up information flow on attention and masking we estimate new overlap scores defined in this papers and based on the so-called ideal binary mask (IBM) [22]. An IBM consists of zeros and ones where ones represent the powerful parts of the target audio signal compared to an interference audio signal. IBMs have been shown to improve human speech intelligibility when applied to noisy speech signals. Subjects have been exposed to the re-synthesized speech signals from the IBM-gated (segregated) signal and they recognized words quite well even for a signal-to-noise-ratio (SNR) as low as -60 dB which corresponds to pure noise ([23], [24]). In addition,the features obtained from IBMs have worked successfully for an automatic speech recognition (ASR) application [8].

The influence of masking is measured as the correlation between the number of times specific words are heard (WOH) and the relative overlap. IBM control parameters are first chosen taking as references previous works on speech intelligibility [24] and ASR [8] as a pre-analysis step. Then, we optimize those parameters to have the most negative correlation. In particular, we analyze local criteria (LC), the window length (winLength) and number of frequency channels (numChan).

The paper is organized as follows; first we discuss the proposed top-down attention model and the experiments designed to investigate the role of attention in the hard cocktail party problem, we present simulation results of the model and experimental results, and finally give our interpretation of the findings in the discussion section.

II. Methods

A. Attention Model

By means of movement and information processing the brain actively selects its input. In the broadest sense attention is the mechanism by which the brain selects relevant input. Bottom-up attention is typically driven by statistical novelty, i.e., we attend to the un-expected, while top-down attention select input that is relevant to a given task. While the latter definition is in broad consensus, there has been remarkably few attempts to formalize top-down attention as a statistical problem. In [5] we model the top-down attention as a decision problem based on incomplete information and analyze which feature to measure next in a classification problem.

Attention is implemented in a statistical model as the selection of additional input features based on an initial subset of features representing the 'gist'. We attend to features that reduce confusion at the models' output level, i.e., features that have high expected information given the 'gist'.

We represent the task by probability distribution over a set of C 'actions' or classes indexed by the discrete variable c, $(c = 1, ..., C)$. Initially, for decision making we have access to the gist, a vectorial observation $\boldsymbol{x}$ with components x_i, $i = 1, ..., I$. The second step concerns an additional measurement z_j which is obtained by attending to a specific channel j, chosen among the set of missing features $\boldsymbol{z}$ with components $z_1, ..., z_J$. The joint probability of the classes and all features observed and missing is $p(c, \boldsymbol{x}, \boldsymbol{z})$. Attention is based on the information available in $\boldsymbol{x}$, giving us the pre-attention posterior probabilities for the task

$$\begin{aligned} p(c|\boldsymbol{x}) &= \int p(c, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \\ &= \frac{\int p(c, \boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}}{\sum_{c=1}^{C} \int p(c, \boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}}. \end{aligned} \tag{1}$$

Using the top down attention mechanism we will select an additional feature $\boldsymbol{z}_j$, which will result in the distribution

$$\begin{aligned} p(c|\boldsymbol{x}, z_j) &= \sum_{c=1}^{C} \int p(c, \boldsymbol{z}|\boldsymbol{x}) \prod_{i \neq j} dz_i \\ &= \frac{\int p(c, \boldsymbol{x}, \boldsymbol{z}) \prod_{i \neq j} dz_i}{\sum_{c=1}^{C} \int p(c, \boldsymbol{x}, \boldsymbol{z}) \prod_{i \neq j} dz_i} \end{aligned} \tag{2}$$

The information value of this choice is given as the difference in confusion (entropy) before and after the attended measurement, which will depend on the particular outcome of the sequential measurement, z_j,

$$\begin{aligned} \Delta S_j(\boldsymbol{x}, z_j) &= \sum_{c=1}^{C} \log p(c|\boldsymbol{x}, z_j) p(c|\boldsymbol{x}, z_j) \\ &- \sum_{c=1}^{C} \int \log p(c, \boldsymbol{z}|\boldsymbol{x}) p(c, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} \end{aligned} \tag{3}$$

As z_j is unknown before attending, we base the choice on the expected gain *expected information gain* of measuring the value of feature j,

$$\begin{aligned} G_j(\boldsymbol{x}) &\equiv \int \Delta S_j(\boldsymbol{x}, z_j) p(z_j|\boldsymbol{x}) dz_j \\ &= \sum_{c=1}^{C} \int \log p(c|\boldsymbol{x}, z_j) p(c, z_j|\boldsymbol{x}) dz_j \\ &\quad - \sum_{c=1}^{C} \int \log p(c, \boldsymbol{z}|\boldsymbol{x}) p(c, \boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z}. \qquad (4) \end{aligned}$$

The Gaussian Discrete mixture model (GDMM) is a generative model of the joint distribution, see e.g., [25]

$$p(c, \boldsymbol{x}, \boldsymbol{z}) = \sum_{k=1}^{K} p(k) p(c|k) p(\boldsymbol{x}, \boldsymbol{z}|k) \qquad (5)$$

where K is the number of components, $p(k)$ are component probabilities, $p(c|k)$ is a $C \times K$ probability table, and $p(\boldsymbol{x}, \boldsymbol{z}|k)$ are K Gaussian pdfs. The GDMM is convenient as both conditioning and marginalization are computationally tractable. We choose a generative representation to allow for modeling of input dependencies which is necessary in order to make inference about missing features. Maximum likelihood parameter estimation in the GDMM leads to a straightforward generalization of expectation maximization algorithm for conventional mixtures.

If we introduce the GDMM in the information gain expression we obtain

$$\begin{aligned} G_j(\boldsymbol{x}) &= \sum_{c=1}^{C} \sum_{k=1}^{K} p(c|k) p(k|\boldsymbol{x}) \times \\ &\qquad \int \log \left[p(c, \boldsymbol{x}, z_j) \right] p(z_j|\boldsymbol{x}, k) dz_j \\ &- \sum_{k=1}^{K} p(k|\boldsymbol{x}) \int \log \left[p(\boldsymbol{x}, z_j) \right] p(z_j|\boldsymbol{x}, k) dz_j \\ &+ \text{const.} \qquad (6) \end{aligned}$$

where $p(c, \boldsymbol{x}, z_j) = \sum_{k=1}^{K} p(k) p(c|k) p(\boldsymbol{x}, z_j|k)$ and $p(\boldsymbol{x}, z_j) = \sum_{k=1}^{K} p(k) p(\boldsymbol{x}, z_j|k)$. Thus, computing G for all I features amounts to computing $Q = I * (C+1) * K$ one-dimensional integrals over Gaussian measures $p(z_j|\boldsymbol{x}, k) = \mathcal{N}(\mu_j(\boldsymbol{x}, k), \sigma_j^2(\boldsymbol{x}, k))$ with

$$\begin{aligned} \mu_j(\boldsymbol{x}, k) &= \mu_{j,k} + \boldsymbol{\Sigma}_{z_j, \boldsymbol{x}, k} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{x}, \boldsymbol{x}, k} (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}, k}) \\ \sigma_j^2(\boldsymbol{x}, k) &= \sigma^2_{j,k} - \boldsymbol{\Sigma}_{z_j, \boldsymbol{x}, k} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{x}, \boldsymbol{x}, k} \boldsymbol{\Sigma}_{\boldsymbol{x}, z_j, k}. \qquad (7) \end{aligned}$$

In these expressions $\mu_{j,k}, \sigma^2_{j,k}$ are the mean and variance of the jth feature in the kth component, while $\boldsymbol{\Sigma}_{a,b,k}$ is the part of the covariance matrix of the k component corresponding to variable sets a, b.

More details and further references are given in [5], where the attention model was validated on four benchmark classification problems and shown to outperform a 'random' attention alternative.

To simulate strong and weak top-down attention we augment the model by smoothing the label-component table $p(c|k) \rightarrow p(c|k, \beta) = p(c|k)^{\beta} / \sum_c p(c|k)^{\beta}$ and by letting the attention selection be stochastic based on the expected gains, i.e., we select attention using the induced probability distribution,

$$P(j) = \frac{\exp(\gamma G_j)}{\sum_j \exp(\gamma G_{j'})} \qquad (8)$$

Task-driven top-down attention as in [5] is obtained when $\beta = 1, \gamma = \infty$. In this work we use $\beta = 0.2, \gamma = 0.33$.

We challenge the strong and weak top-down attention models by a simulated cocktail party by confounding the input of test data for a $C = 2$ environment. In particular we mix for each pattern a fraction f (mixing fraction) of the input features with a randomly chosen input feature vector from the opposite class (confounder):

$$overlapped\ signal = (1 - f) * input\ signal + f * confounder$$

The $C = 2$ simulated decision problem is based on a four component Gaussian mixture, the resulting configuration is first established in two dimensions and resembles the well-known XOR-problem, hence can not be separated linearly. The two dimensional input space is augmented by six noise dimensions $SNR \approx 1$, so that the total input dimension is eight. In the attention experiments one signal dimension and one noise dimension is provided as 'gist'.

B. Behavioral Experiments

While it is not possible to directly read out the informations flows in the human brain while solving a difficult speech separation task, some insight can be obtained by observing the macroscopic behavior. Here we design a behavioral experiment inspired by the pioneering work of Cherry [1]. Basically, we design a hard cocktail party problem by reducing conventional auditory cues as described above, leaving only high-level cognitive cues such as semantic and context representation in narratives. Cherry alluded to these high-level representations as what he called word *transition probabilities*. Our hypothesis is that these representations precisely are subject to top-down attention and should be task dependent, while the more basic cues could be more automatic and operate beyond conscious control.

Subjects are presented with two different narratives combined in a mono audio file with a headphone. The stories are generated by a speech synthesizer (TTS project at AT&T Labs Research [17]), using the same virtual speaker.

We recruited twelve participants among master students, PhD students and post-docs from the Technical University of Denmark. Two participants were not able to accomplish any of the experiments and were excluded.

We perform two different types of experiments which we will refer to as *undirected attention* and *directed attention* experiments. In the first case the listener is free to follow either narrative. In the directed attention case, participants are asked to follow one narrative story at their own discretion. At the end of an audio presentation, a list of terms is presented and they are asked to check which terms they have heard. The list contains words which are in the narratives and words which are not, but are related to the content.

Cherry made his participants listen as many times as the they wanted; we perform three different trials (4 people for each trial). In the first and in the second, we make them listen just once and then we ask for the words. They repeat the same process three times. In the third, we make them listen twice before presented with the term lists.

The words in the list can appear various numbers of times in the narratives, but we balance this number, in total, for both tracks. In particular, the total number of occurrences of all the words we ask for is the same in both stories. Moreover, we aimed to balance the frequency of each word in the list; which means, for example, that if there are two words appearing in the list, respectively, three and five times in the first story, there are also two words appearing three and five times in the second one. The list of words contains 48 words: 24

of these are truly present in the audio signal and each story contains half of them.

We use two different narratives for each experiment, making small changes (removing or adding sentences or words, switching the order of some sentences or words) to the original texts. This is necessary to have stories with the same length and in which pauses are synchronized as much as possible, to avoid the so called ‘listening in the gaps’ effect, described by Bregman [9] and by Bronkhorst [12].

C. Audio Analysis

The basis for our audio analysis is the *ideal binary masks* (IBMs) which we use to measure the cross channel interference in terms of temporal and spectral overlap. It is obtained by comparing the spectrogram of a target sound signal to that of the interference signal and to keep only the strong time-frequency regions of it. More specifically, its value is one when the target is stronger than the interference for a local criterion (LC), and zero elsewhere. The time-frequency (T-F) representation is obtained by using the model of the human cochlea as the basis for data representation [26]. If $T(t, f)$ and $I(t, f)$ denote the target and interference time-frequency magnitude, then the IBM is defined as

$$\mathrm{IBM}(t,f) = \begin{cases} 1, & \text{if } T(t,f) - I(t,f) > LC \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

In Figure 1, we show an example of an IBM obtained with a sample sound from one of the stories (the sound relative to the word ‘navigate’) compared to a speech shaped noise (SSN) as the interference signal. The most energetic parts of the target signal are kept. We measure the spectral and temporal overlap between two sound signals, specifically between a word in one of the narratives pronounced by the speech synthesizer and the corresponding part in the second story. We define the temporal overlap between them as a percentage of the whole duration without silence in the time domain. We use IBMs of the sound signals as mentioned before and we compress both IBMs over frequency. For a time slot, we assign one if there is at least one one on the mask, otherwise zero where zeros are considered as silence. Then, the temporal overlap is simply the overlap of ones on the masks (see Figure 2).

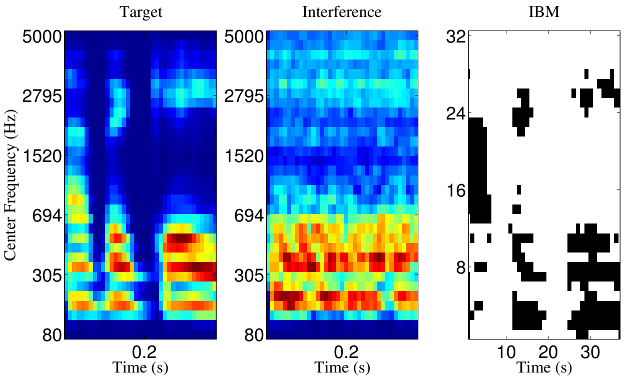

Fig. 1. The spectrograms of a target sound signal, the interference signal (SSN) and the resultant IBM (SNR=0 dB, LC=-4 dB, windows length=20 ms (50% overlap), frequency channels=32, frequency bins are not equally spaced, gammatone filtering is used, 1 = black 0 = white)

The spectral overlap is defined similarly based on co-occurrence of signals in the time-frequency bins. Once we have IBMs for both sound signals, we simply compute the percentage of the overlapping ones on both masks over the whole time-frequency bins (see Figure 2).

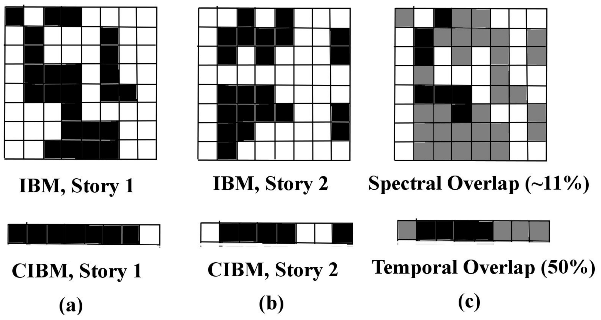

Fig. 2. The illustrations of temporal and spectral overlap definitions, the bins represent time-frequency regions of an IBM (frequency bins are not equally spaced, gammatone filtering is used). Only black regions represent overlapped parts on (c).

Based on the temporal and spectral overlaps for words in both stories, we explore the correlation between the overlaps and the number of times the words were heard by the subjects for both directed and undirected cases.

The overlap values depend on parameters that change the resultant IBMs, while the number of times the words were heard is fixed. In particular, for each word, the number we consider in the analysis is computed averaging out the total number of times the word was heard (in all the experiments, in all the trials, by all the subjects) against the number of times the same word occurs in the stories. IBM parameter values are first chosen taking as references previous works on speech intelligibility [24] and on ASR [8] as a pre-analysis step. In [24], subjects listened to IBM-gated (multiplying the spectrogram of a noisy-speech with IBM of it, and resynthesizing in time domain) and for a range of IBM parameter values, the best speech intelligibility results(recognizing which word is pronounced) are obtained. In [8], the best performance for an ASR system is obtained again with same range of values. However, referenced to those studies, even if those parameters are expected to result in overlaps closer to what humans perceive, they are not necessarily optimal to investigate the correlation between overlap and word detection rates. Therefore we optimize the IBM parameters including the local criteria (LC) with fixed SNR, the windows length and the number of frequency channels to gain the most negative correlation. We keep other two parameters constant while optimizing one. With the optimized values, we apply a permutation test with 10000 resamples, at 5% significance level, to validate the results.

III. Experimental Results

First, we set up a simulation experiment with the top-down attention model emulating the Cherry experiment, as described above. For a range of mixing fractions $f \in [0.0 \;\; 0.2]$ we measure the resulting error rates for the models using the attention mechanism to select an additional feature among the six remaining after having provided the 2-dimensional gist. The strong and weak top-down attention response is shown in figure 3. The rates are represented as relative excess errors: $[E(f) - E(0)]/(B - E(0))$, where $B = 0.5$ is the baseline error rate. The error rate of the top-down attention model is $E_{\mathrm{DIR}}(0) = 0.08$ while the error rate of the weak attention is $E_{\mathrm{UNDIR}}(0) = 0.23$. The experiment indicates that strong top-down attention model (DIR) is less sensitive to the confounding mixture than the weak attention model, hence it will make more informed decisions in the cocktail party.

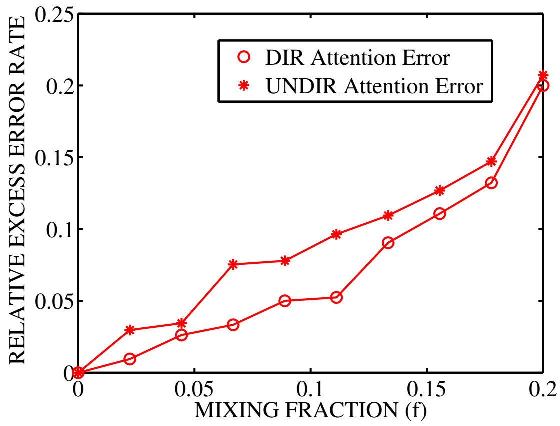

Fig. 3. The resulting error rates for the models using the attention mechanism to select an additional feature among the six remaining after have provided the 2-dimensional gist for a range of mixing fractions $f \in [0.0\ 0.2]$.

The behavioural experiments are carried out using a GUI implemented in Java, while the results are analysed using MATLAB. The word boundaries are determined manually to be more precise (The limited number of words enables us to do so). The sampling frequency of the audio signals is 16 kHz. We use gammatone filters which is a commonly used model for auditory filters in the auditory system to obtain IBMs.

Figures 4 and 5 show the temporal and spectral overlaps for each word, for UNDIR and DIR cases respectively, using non-optimized parameters from [24], [8]. We observe that the correlation between overlaps (temporal and spectral) and rate of heard words are -0.35 and -0.31 respectively. While, for DIR case, we find positive correlations of 0.23 for temporal and 0.34 for spectral. We next optimize the three IBM parameters, LC, WinLength and NumChan, to produce the most negative correlation between overlaps and words detected. The resulting figures show that optimal LC values are around -10dB for all cases except for the spectral overlap in the UNDIR case, which is -14dB (see Figure 6). We also conclude that 20ms is the optimal value for the windows length for all cases (see Figure 7). We see that for the spectral overlap in the DIR case, the correlation values for WinLength greater than 20ms are not present. This is due to the fact that with the high optimal value found previously, the resultant IBMs were all zeros (we did not try to play with the values, because it is already hard to find significant results for DIR case). Finally, we observe that the optimal values for number of frequency channels is 16 and 32 (see Figure 8). Using optimal IBM parameters for each case

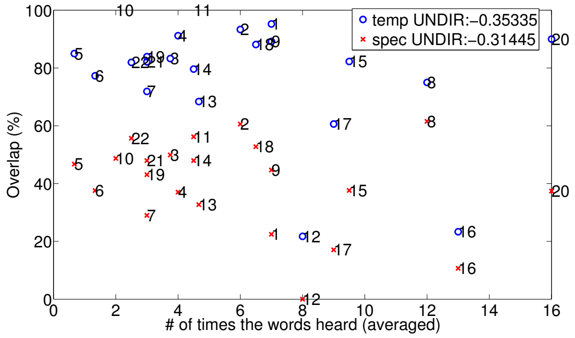

Fig. 4. Temporal and spectral overlap versus the averaged number of times the words heard (WOH) for UNDIR case, and the correlation between them shown on the legend

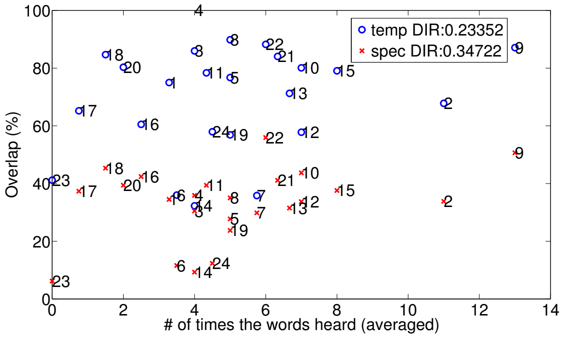

Fig. 5. Temporal and spectral overlap versus the averaged number of times the words (WOH) heard for DIR case, and the correlation between them shown on the legend

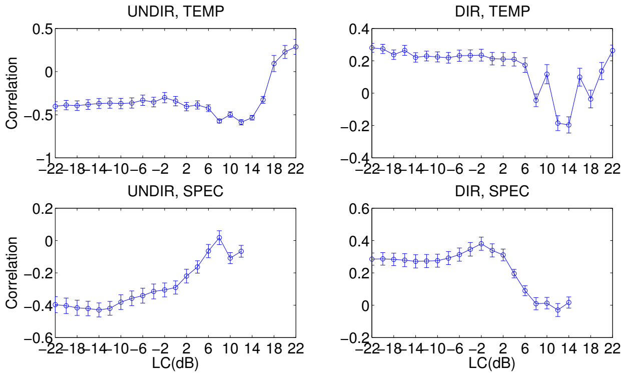

Fig. 6. Correlation for different LC values, WinLength = 20ms and NumChan = 32. Left to right: Undirected and Directed. Top to bottom: Temporal and Spectral.

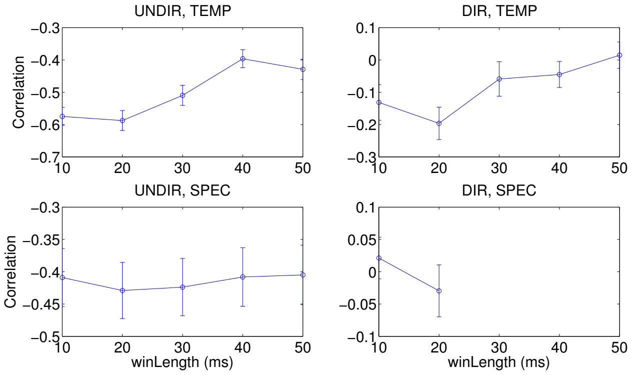

Fig. 7. Correlation for different WinLength values, optimal LC for each case and NumChan = 32. Left to right: Undirected and Directed. Top to bottom: Temporal and Spectral.)

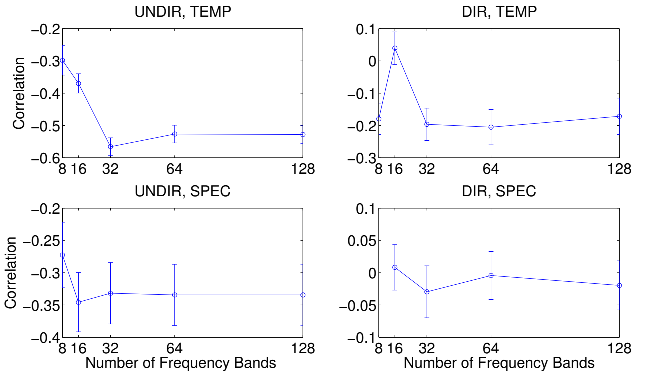

Fig. 8. Correlation for different NumChan values, optimal LC and WinLength for each case. Left to right: Undirected and Directed. Top to bottom: Temporal and Spectral.)

(UNDIR,DIR, temporal and spectral) we obtain similar results. The correlations between overlaps (temporal and spectral) and WOH are -0.59 and -0.43 (more negative than not-optimized case) respectively for UNDIR case. However, for DIR case, they are -0.20 for temporal and -0.03 for spectral. Even if the results for DIR case also are more negative than not-optimized case , they are evidently less than those of UNDIR case (almost no correlation for DIR spectral case). Finaly, we apply the permutation test to these data, as mentioned in the section II-C. In both spectral and temporal overlaps, for UNDIR experiments, under the 5% significance level, the null hypothesis that the data is uncorrelated is rejected (spectral = 3.1% and temporal = 0.07%). In the DIR experiments the null hypothesis is accepted, indicating the influence of masking is compensated by a more detailed model.

A. *Discussion and Conclusion*

Based on our recent top-down attention model we can simulate the cocktail party effect. We found that the top-down attention model showed less sensitivity to the amount of the confounding overlap, than the weak attention model. This indicates that the top-down mechanism can assist to compensate for structured noise.

In the 'hard cocktail party' behavioral experiment we found significant negative correlations between overlaps of two concurrent sounds and speech intelligibility for the data collected in the undirected attention experiments (UNDIR, no task). While in the directed attention experiments (DIR, task-driven) we accepted the no-correlation null hypothesis, even after careful optimization for correlation, a finding well-aligned with the simulation result.

We conclude that the relation between energetic masking and speech intelligibility is modulated by the presence of a task, hence top-down controlled attention. Based on our top-down attention model we expect this to be a special case of a more general phenomenon, namely that the top-down knowledge can enhance pattern recognition by compensating for noise and the presence of confounders.

References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and two ears," *Journal pf Acoustic Society of America*, vol. 25, pp. 975–979, 1953.

[2] M. McKeown, L. K. Hansen, and T. J. Sejnowski, "Independent component analysis for fmri: What is signal and what is noise?" *Current Opinion in Neurobiology*, vol. 13, no. 5, pp. 620–629, 2003.

[3] A. J. Viterbi, *CDMA: principles of spread spectrum communication*. Redwood City, CA, USA: Addison Wesley Longman Publishing Co., Inc., 1995.

[4] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis*. J. Wiley, 2001.

[5] L. Hansen, S. Karadogan, and L. Marchegiani, "What to measure next to improve decision making? On top-down task driven feature saliency," in *SSCI, IEEE Symposium on Computational Intelligence, Paris, France. CCMB Cognitive Algorithms, Mind, and Brain*, 2011, pp. 86–87.

[6] S. Karadogan, L. Marchegiani, J. Larsen, and L. Hansen, "Top-down attention with features missing at random," in *IEEE International Workshop on Machine Learning For Signal Processing*, 2011, submitted.

[7] S. Choi, H. Hong, H. Glotin, and F. Berthommier, "Multichannel signal separation for cocktail party speech recognition: A dynamic recurrent network," *Neurocomputing*, vol. 49, no. 1-4, pp. 299–314, 2002.

[8] S. Karadogan, J. Larsen, M. Pedersen, and J. Boldt, "Robust isolated speech recognition using binary masks," *European Signal Processing Conference, EUSIPCO*, 2010.

[9] A. Bregman, *Auditory scene analysis: The perceptual organization of sound*. The MIT Press, 1994.

[10] B. Moore and H. Gockel, "Factors influencing sequential stream segregation," *Acta Acustica United with Acustica*, vol. 88, no. 3, pp. 320–333, 2002.

[11] R. Drullman and A. Bronkhorst, "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," *The Journal of the Acoustical Society of America*, vol. 107, p. 2224, 2000.

[12] A. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

[13] M. Bee and C. Micheyl, "The cocktail party problem: What is it? how can it be solved? and why should animal behaviorists study it?." *Journal of Comparative Psychology*, vol. 122, no. 3, p. 235, 2008.

[14] R. Cusack, J. Decks, G. Aikman, and R. Carlyon, "Effects of location, frequency region, and time course of selective attention on auditory scene analysis." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 30, no. 4, p. 643, 2004.

[15] R. Carlyon, R. Cusack, J. Foxton, and I. Robertson, "Effects of attention and unilateral neglect on auditory stream segregation." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 27, no. 1, p. 115, 2001.

[16] D. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *The Journal of the Acoustical Society of America*, vol. 109, p. 1101, 2001.

[17] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The at&t next-gen tts system," in *Joint Meeting of ASA, EAA, and DAGA*. Citeseer, 1999, pp. 18–24.

[18] A. Syrdal and Y. Kim, "Dialog speech acts and prosody: Considerations for tts," in *Proceedings of Speech Prosody*, 2008, pp. 661–665.

[19] M. Jilka, A. Syrdal, A. Conkie, and D. Kapilow, "Effects on tts quality of methods of realizing natural prosodic variations," in *Proc. ICPhS*, 2003.

[20] D. Phillips, *Preparation Course for the TOEFL: Next Generation IBT*. Longman, 2006.

[21] A. Oliva and A. Torralba, "Building the gist of a scene: the role of global image features in recognition," in *Progress in Brain Research*, 2006, p. 2006.

[22] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech separation by humans and machines*, pp. 181–197, 2005.

[23] D. Wang, U. Kjems, M. Pedersen, J. Boldt, and T. Lunner, "Speech perception of noise with binary gains," *The Journal of the Acoustical Society of America*, vol. 124, pp. 2303–2307, 2008.

[24] U. Kjems, J. Boldt, M. Pedersen, T. Lunner, and D. Wang, "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *The Journal of the Acoustical Society of America*, pp. 1415–1426, 2009.

[25] L. K. Hansen, S. Sigurdsson, T. Kolenda, F. Å. Nielsen, U. Kjems, and J. Larsen, "Modeling text with generalizable gaussian mixtures," in *ICASSP International conference on acoustics, speech and signal processing*, vol. 4, 2000, pp. 3494–3497.

[26] R. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82.*, vol. 7, 1982, pp. 1282–1285.