MODELING THE PERCEPTION OF MULTITALKER SPEECH

Soundararajan Srinivasan

Biomedical Engineering Center
The Ohio State University
Columbus, OH 43210, USA
srinivasan.36@osu.edu

DeLiang Wang

Department of Computer Science &
Engineering and Center for Cognitive Science
The Ohio State University
Columbus, OH 43210, USA
dwang@cse.ohio-state.edu

Abstract

Listeners' ability to understand a target speaker in the presence of one or more simultaneous competing speakers is subject to two types of masking: Energetic and informational. Energetic masking occurs when target and interfering signals overlap in time and frequency resulting in portions of target becoming inaudible. Informational masking occurs when the listener is unable to segregate the target from interference, while both are audible. We present a model of multitalker speech perception that accounts for both types of masking. Human perception in the presence of energetic masking is modeled using a speech recognizer that treats the masked time-frequency units of target as missing data. The effects of informational masking on the recognizer are modeled using the output of a speech segregation system. On a systematic evaluation, the performance of the proposed model is in broad agreement with perceptual results.

1. Introduction

In everyday listening conditions, the acoustic input reaching our ears is often a mixture of multiple sound sources. In such situations, the human ability to perceive a target source is degraded by the effects of masking, which is defined as the increase in the audibility threshold of the target caused by the presence of other sources in the environment [1]. In particular, our ability to attend to and understand a target speaker, in the presence of simultaneous competing talkers, is affected by at least two types of masking: Energetic and informational. Energetic masking refers to the phenomenon in which a stronger signal masks a weaker one within a critical band [1]. Informational masking refers to the perceptual degradation caused by listeners' inability to segregate audible portions of target from interference [2]. In this paper, we propose a model for recognizing the contents of a target speech signal subject to both types of masking under monaural conditions. Apart from modeling related psychophysical data, the proposed model can also serve as a paradigm for automatic speech recognition in multitalker situations.

Spectro-temporal overlap between target and interference is a prime cause of energetic masking. Portions of target subject to energetic masking become inaudible at the periphery of the auditory system and are unavailable for subsequent processing. A missing-data speech recognizer [3] is therefore used to model listener perception under energetic masking conditions. When target speech is contaminated by additive interferences, some time-frequency (T-F) regions will contain predominantly target energy (reliable) and the rest are subject to energetic masking by interference. The missing data method will be used to treat the latter T-F regions as missing or unreliable during recognition. The missing data recognizer requires a binary T-F mask that provides information about which T-F regions, of the mixture signal, are reliable and which are unreliable. The task of generating such a mask is akin to the task of segregating the target from the mixture. The process by which the auditory system is able to organize the acoustic input into components that correspond to individual sources in the input is known as auditory scene analysis (ASA) [4]. Therefore, informational masking is intertwined with ASA. Hence, we adapt a monaural computational auditory scene analysis (CASA) system [5] to estimate a binary mask that selects T-F regions of the mixture where target dominates interference. The similarities between target and interference characteristics affect the performance of the CASA system and therefore contribute to informational masking in our model. Cooke also used a missing-data recognizer to model listeners' perception in babble noise [6]. His model used *a priori* knowledge of speech-dominant T-F regions.

The model proposed here can also serve as an architecture for robust speech recognition in the presence of multiple interfering speech sources. It is well known that the performance of automatic speech recognizers (ASRs) degrades rapidly in the presence of interfering sound sources [7, 8]. Speech recognizers are typically trained in an environment containing a single speech source and face a problem of mismatch when used in conditions where target speech occurs simultaneously with other sound sources. To mitigate the effect of this mismatch on recognition, "noisy" speech is typically preprocessed by speech separation systems. However, in many realistic applications, the output of typical speech segregation algorithms contains distortions in segregated speech not seen during ASR training. These distortions cause substantial degradation in recognition performance [3]. Target speech distortions, even in the case of perfect segregation is similar to the energetic masking. Additionally, the task of target speech segregation itself is directly related to the informational masking problem as mentioned above. Hence a model that accounts for both types of masking should should also improve the robustness of ASRs.

The rest of the paper is organized as follows. A recent study examined the degradation in listeners' perception caused by energetic and informational masking using a binary masking procedure [9, 10]. This study is briefly reviewed in the next section. Section 3 contains a detailed presentation of the proposed model. The model has been systematically evaluated on the same task as in [9, 10]. The evaluation results and a comparison with listeners' performance is presented in Section 4. Finally, conclusions and future work are given in Section 5.

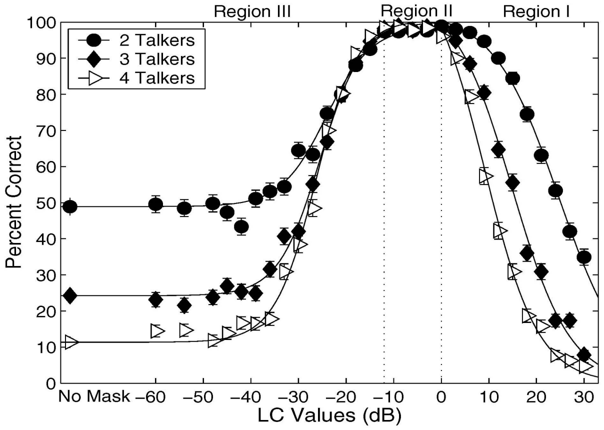

Figure 1: *Percentage of trials in which the listeners identified the keywords in the target phrase correctly (from [9, 10]). The error bars represent* 95% *confidence intervals. The figure shows the effect of number of competing talkers on listener performance.*

2. Energetic and informational masking in multitalker speech perception

A recent study [9, 10] used ideal binary T-F masks to isolate the effects of energetic and informational masking on the intelligibility of a target speech signal in the presence of one or more competing speech signals. An ideal binary mask is obtained *a priori*, from premixing target and interference. An unit in the ideal binary mask is assigned a value 1 if the signal to noise ratio (SNR) calculated from the corresponding T-F units of target and interference exceeds a predefined local SNR criterion (LC) value; it is labeled 0 otherwise. To generate this mask, target and interfering signals are first analyzed using a 128 channel gammatone filterbank whose center frequencies are quasi-logarithmically spaced from 80 Hz to 5 kHz [5]. The energy at the output of each filter is calculated every frame for both target and interference. Each frame is 20 ms long with 10 ms frame shift. The ideal binary mask is used to resynthesize a signal by retaining only those T-F units in the mixture where the local SNR exceeds the specified LC value [9]. The resynthesized signal is then used in a series of experiments to study effects of number of interfering talkers and their sex on the two types of masking.

The speech corpus used in [9, 10] is the CRM corpus [11]. This corpus consists of utterances from 4 male and 4 female speakers produced according to the grammar, “READY ($call-sign) GO TO ($color) ($digit) NOW”. There are 8 call signs, 4 colors and 8 numbers [11] and the target utterance always contains the call sign “BARON”; e.g., “READY BARON GO TO RED ONE NOW”. The interference utterance consists of a call sign, a color and a number different from that of the target. The task for the listener is to identify the color and number in the target phrase. The target-to-masker ratio (TMR) was fixed at 0 dB [9]. While TMR is used to refer to the ratio of target to each interference in the mixture, SNR is used to refer to the ratio of target to combined interference energy [2]. Fig. 1 shows the effects of varying the number of competing talkers on the correct identification of both color and number in the target phrase as a function of the LC value [9, 10]. The information lost in the target signal due to energetic masking is proposed to be the same as the one resulting from the resynthesis using an ideal binary mask with a LC value of 0 dB [9, 10]. Note that this mask removes all interference dominant T-F units. The signal resynthesized from an ideal binary mask with a positive LC value of δ dB is also shown to cause energetic masking equivalent to that caused by a mixture with TMR of $-\delta$ dB. Further, decreasing the LC value below 0 dB allows for the progressive introduction of interference energy in the resynthesized signal. This leads to an increase in the level of informational masking [10]. Energetic masking effects therefore dominate listeners’ performance in Regions I and II, while informational masking effects dominate in Region III in Fig. 1 [9, 10].

3. A computational model

The gradual degradation in listener performance when LC value decreases in Region III (Fig. 1), can be attributed to ASA. Listeners are able to use the characteristics of the interfering talkers to segregate the target speaker with varying degrees of success. The performance limitations of ASA are manifest as informational masking. Additionally, Region II and parts of Region I demonstrate the robust performance of listeners in the presence of energetic masking as isolated by the use of ideal binary masks. The above observations motivate our model for multitalker speech perception as shown in Fig. 2. A monaural CASA system [5] is adapted to segregate a target from interfering sources. Perfect segregation by the CASA system will result in a binary mask that can be used to remove all interference dominant T-F regions from the mixture [5], in other words a binary mask that isolates energetic masking of target [9, 10]. Hence, listener perception in the presence of energetic masking is modeled using a missing-data ASR that treats masked data as missing target-data. The similarities between target and interference characteristics cause deviations in the estimation of the binary mask by the CASA system. This contributes to informational masking in our model.

The input to the model is a mixture of target and interference, sampled at 20 kHz. We use the same auditory filterbank decomposition of the input signal as used in [9, 10] (see Section 2). The output is used to generate feature vectors for recognition and as input to a monaural CASA system. Under monaural conditions, the human auditory system can segregate a target speech source from various interference using several cues, including differences in pitch and onsets [4]. Pitch has been successfully used for segregation of voiced speech in several CASA systems [12, 5]. Hence, we adapt the speech separation system in [5] to segregate target from interference. This system is chosen as it shows robust performance when tested with a variety of intrusions. The system is based on two main stages: 1) segmentation and 2) grouping. In segmentation, the input signal is decomposed into a collection of contiguous T-F units that are dominated by one sound source. During grouping, those segments that are likely to belong to the same source are grouped together. In the low-frequency range, the system generates segments based on temporal continuity and cross-channel correlation, and groups them based on periodicity similarity. For high-frequencies, the signal envelope fluctuates at the pitch rate and amplitude modulation rates are used for grouping [5]. To illustrate the potential of the proposed approach for segregation, pitch tracks and pitch strengths are derived *a priori* from premixing target and interference signals using Praat [13]. Robust multipitch tracking of more than two sources is a challenging problem currently [14]. Note that the system in [5] utilizes only the target pitch contour for grouping. However, psychoacoustic evidence suggests that

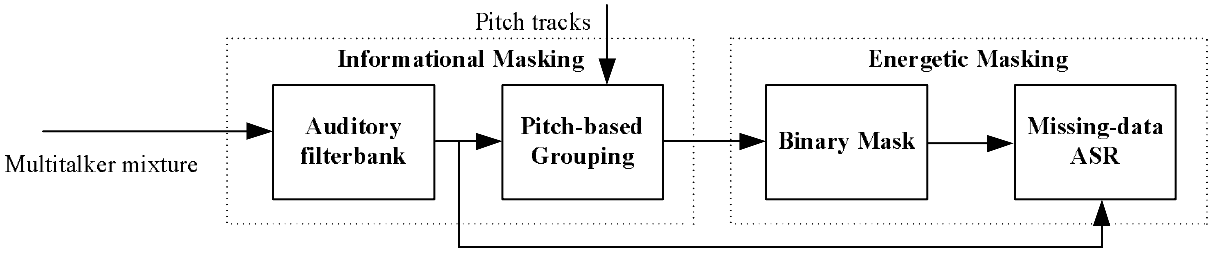

Figure 2: *Block Diagram of the proposed model. The input mixture signal is analyzed by an auditory filterbank in successive time frames. The output is fed to a monaural CASA system that uses pitch tracks of the individual sources in the mixture to produce a binary mask that selects T-F regions in the mixture where target dominates interference. This mask is used by the missing-data recognizer to decode the input.*

listeners are able to utilize the pitch information of interference sources too [4]. Therefore, we adapt the system in [5] to group segments based on the dominant pitch at a given time-frame. If the dominant pitch at a particular time belongs to interference, we discard the grouped T-F units in that frame. On the other hand, if the dominant pitch belongs to the target, the grouped T-F units are retained. We do not current process the unvoiced regions. The output of the CASA system is therefore an estimate of a binary mask that labels the target-dominant regions in the mixture as reliable (1) and the rest as unreliable (0). This mask is then used by the missing-data ASR to recognize target speech.

The input to the missing-data recognizer is the instantaneous Hilbert envelope at the output of each gammatone filter, smoothed using a first-order filter with 8 ms time constant and log compressed as suggested in [3]. The missing data recognizer [3] makes use of spectro-temporal redundancy in speech to recognize a "noisy" signal based on its target dominant T-F units. Given an observed speech vector Y, the word recognition problem is to maximize the posterior $P(\omega_i|Y)$, where ω_i is a valid word sequence according to the grammar for the recognition task. When parts of Y are masked by interference, Y can be partitioned into its reliable and unreliable constituents as Y_r and Y_u. In the marginalization method, the posterior probability using only the reliable constituents is computed by integrating over the unreliable ones [3]. If Y represents spectral magnitude and sound sources are additive, the unreliable parts can be constrained as $0 \leq Y_u^2 \leq Y^2$. This bounded marginalization method is shown in [3] to have a better recognition score than the simple marginalization method, and is hence used in all our experiments.

4. Experimental results

To facilitate a comparison with listeners' performance from [9, 10], we have evaluated our model also using the CRM corpus [11]. Sixteen (ready, baron, goto, blue, green, red, white, now and the numbers 1-8) speaker-independent word-level HMM models are trained. All models have 10 states, whose output distribution is modeled as a mixture of 2 Gaussians. The models are trained using 224 utterances from 3 male talkers (Talkers 1-3) and 4 female talkers (Talkers 4-7) in this database that contain the call sign "BARON". The testing data consists of utterances from a male talker (Talker 0), not utilized during training and containing the call sign "BARON". Similar to the experiments in [9, 10], 1, 2 and 3 utterances from the same talker are added to target as interference. The TMR is 0 dB. Interference utterances contain call signs, numbers and colors different from target. Each testing condition comprises of 256 utterances. A HMM toolkit, HTK [15] is used for training. During testing, the decoder is modified to incorporate the missing data methods. The task is recognition of both the color and number in the target utterance. Recognition performance on the target-only input is 100%. The baseline recognition performances on the mixture data with 2-, 3- and 4-talkers are 15.6%, 6.2% and 4.3% respectively.

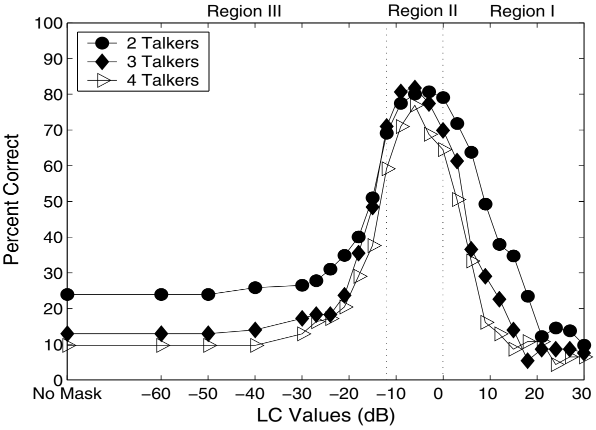

Figure 3: *Percentage of utterances in which the model identified the keywords in the target utterance correctly. The figure shows the model's performance with respect to the LC value under various multitalker conditions.*

To compare with the listener performance from [9, 10], the mixture-data is resynthesized using ideal binary masks with varying LC values (see Section 2). The resynthesized signal is used as input to the proposed model. Fig. 3 shows the performance of our model as a function of the LC value. While the absolute recognition rates are lower, the model is able simulate the general pattern of listeners' performance seen in Fig. 1. As in [9, 10], informational masking can be seen to dominate the recognition performance. "No Mask" represents the control condition in which the mixture signal is used directly as input. The model performance with this input shows a small but consistent improvement over the baseline recognition performance for each of the three conditions. Since interference utterances come from the same speaker, target and interference pitch contours are close to each other. This is the cause for the limited success of pitch-based segregation and hence missing-data ASR.

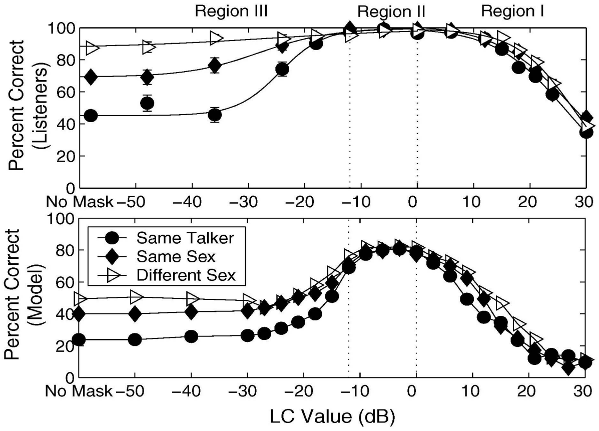

Figure 4: *The effect of voice characteristics on listener and model performance with 2-talkers as a function of the LC value. The top panel shows the performance of listeners in correctly identifying the keywords in the target utterance (from [9, 10]). The bottom panel shows the proposed model's performance on the same task.*

Fig. 4 shows the results of a second experiment in which the effects of voice characteristics of an interfering talker are examined. The top-panel shows listener performance under three different interference conditions: Same talker, a different talker of the same sex and a different sex talker [9, 10]. The bottom-panel shows the performance of our model. For the "Same Sex" condition, interference utterances are chosen from Talkers 1-3. For the "Different Sex" condition, interference utterances are chosen from Talkers 4-7. The baseline recognition performance for the former condition is 17.9%, while for the latter condition is 21.2%. The pitch-based grouping component of the model is now able to utilize the larger differences between target and interference pitch contours to produce better segregation results. This improves the ASR performance. Note that for both model and listeners, differences in voice characteristics between target and interference have only a negligible effect on the performance in Regions I and II. In both cases, the peak performance exhibits a plateau for LC values in the range -9 dB to 0 dB.

5. Conclusions

We have presented a model for monaural multitalker speech perception that is able to account for the effects of both energetic and informational masking. Using this model we have simulated several aspects of listeners' performance including the differential effects of energetic and informational masking on multitalker perception. We have also shown that differences between target and interference pitch ranges contribute to a reduction in informational masking by improving target speech segregation. The use of other ASA cues including common onset should help further enhance segregation, especially for unvoiced speech [16]. Note that as in [9, 10], we have only addressed simultaneous masking in the present study. The proposed model provides significant improvement over baseline recognition performance and hence shows potential for robust speech recognition. Future work will attempt to improve the performance of the missing-data ASR to help bridge the gap between our model performance and that of listeners.

ACKNOWLEDGMENTS. This research was supported in part by an AFOSR grant (FA9550-04-1-0117) and an AFRL grant via Veridian. We thank P. S. Chang for helpful suggestions.

6. References

[1] B. C. J. Moore, *An introduction to the Psychology of Hearing*, 4th ed. San Diego, CA: Academic Press, 2003.

[2] D. S. Brungart, B. D. Simpson, M. A. Ericson, and K. R. Scott, "Informational and energetic masking effects in the perception of multiple simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 110, pp. 2527–2538, 2001.

[3] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, pp. 267–285, 2001.

[4] A. S. Bregman, *Auditory scene analysis*. Cambridge, MA: The MIT Press, 1990.

[5] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. on Neural Networks*, vol. 15, pp. 1135–1150, 2004.

[6] M. Cooke, "A glimpsing model of speech perception," in *Proc. ICPhS '03*, 2003, pp. 1425–1428.

[7] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Comm.*, vol. 16, pp. 261–291, 1995.

[8] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, pp. 2236–2252, 2003.

[9] P. S. Chang, "Exploration of behavioral, physiological, and computational approaches to auditory scene analysis," Master's thesis, Department of Computer Science & Engineering, The Ohio State University, 2004. Available at http://www.cse.ohio-state.edu/pnl/theses/Chang_MSThesis04.pdf

[10] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with an ideal binary mask," Submitted for journal publication.

[11] R. S. Bolia, W. T. Nelson, and M. A. Ericson, "A speech corpus for multitalker communications research," *J. Acoust. Soc. Am.*, vol. 107, pp. 1065–1066, 2000.

[12] G. J. Brown and M. P. Cooke, "Computational auditory scene analysis," *Comp. Speech and Lang.*, vol. 8, pp. 297–336, 1994.

[13] P. Boersma and D. Weenink, "Praat: doing Phonetics by Computer, Version 4.0.26," 2002. Available at http://www.fon.hum.uva.nl/praat

[14] G. J. Brown and D. L. Wang, "Separation of speech by computational auditory scene analysis," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. NY: Springer, 2005, pp. 371–402.

[15] S. Young, D. Kershaw, J. Odell, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.0)*. Microsoft Corporation, 2000.

[16] G. Hu and D. L. Wang, "Separation of fricatives and affricates," in *Proc. ICASSP '05*, 2005, pp. I. 1101–1104.