Speech perception of noise with binary gains

DeLiang Wang[a)]
Department of Computer Science & Engineering, and Center for Cognitive Science, The Ohio State University, Columbus, Ohio 43210

Ulrik Kjems, Michael S. Pedersen, and Jesper B. Boldt
Oticon A/S, Kongebakken 9, DK-2765 Smørum, Denmark

Thomas Lunner
Oticon Research Centre Eriksholm, Kongevejen 243, DK-3070 Snekkersten, Denmark and Department of Clinical and Experimental Medicine, and Technical Audiology, Linköping University, S-58183 Linköping, Sweden

(Received 9 December 2007; revised 7 July 2008; accepted 8 July 2008)

For a given mixture of speech and noise, an ideal binary time-frequency mask is constructed by comparing speech energy and noise energy within local time-frequency units. It is observed that listeners achieve nearly perfect speech recognition from gated noise with binary gains prescribed by the ideal binary mask. Only 16 filter channels and a frame rate of 100 Hz are sufficient for high intelligibility. The results show that, despite a dramatic reduction of speech information, a pattern of binary gains provides an adequate basis for speech perception.
 [DOI: 10.1121/1.2967865]

PACS number(s): 43.71.Gv, 43.66.Dc [KWG] Pages: 2303–2307

I. INTRODUCTION

Human speech recognition shows remarkable robustness in a variety of listening conditions, including competing talkers, environmental sounds, and background noise. Understanding how speech is recognized under these conditions is important not only for auditory perception but also for automatic speech recognition, where robustness to acoustic interference remains elusive (Lippmann, 1997; Allen, 2005). Related research in computational auditory scene analysis (CASA) and blind source separation makes use of a binary time–frequency ($T-F$) masking technique (Roman *et al.*, 2003; Hu and Wang, 2004; Yilmaz and Rickard, 2004). Time–frequency masking operates on a $T-F$ representation or decomposition of the input into a two-dimensional matrix of $T-F$ units. Such a representation can be readily generated by passing the input signal through a filterbank and then time windowing the response of each filter. Then binary masking as a means of separation amounts to identifying a binary mask where 1 indicates that the acoustic energy in the corresponding $T-F$ unit is retained and 0 indicates that the energy is removed. In other words, binary masking applies a pattern of binary gains to the mixture signal. It should be noted that the term "masking" here means weighting the mixture, which is different from the same term used in psychoacoustics where it means blocking the target sound by acoustic interference.

Among $T-F$ masks, the so-called ideal binary mask (IBM) has been suggested to be a goal of CASA (Wang, 2005). The IBM is a matrix where 1 indicates that the signal-to-noise ratio (SNR) within the corresponding $T-F$ unit exceeds a threshold LC (local SNR criterion) and 0 otherwise. The mask is "ideal" because its construction requires that speech and noise be available before they are mixed, and the mask possesses certain optimality in terms of overall SNR gain when LC is set to 0 dB (Li and Wang, 2008).

Recent studies in speech perception show that applying IBM to noisy speech leads to large speech intelligibility improvements (Brungart *et al.*, 2006; Anzalone *et al.*, 2006; Li and Loizou, 2008). In particular, Brungart *et al.* (2006) and Li and Loizou (2008) have shown that, with fixed levels of input SNR (between −10 and 0 dB), a range of LC values produces nearly 100% correct scores. The large intelligibility gain has been attributed to ideal segregation (or detection) that directs the listener's attention to the $T-F$ regions of noisy speech where the target speech is relatively strong. This explanation emphasizes the importance of the target signal contained in the $T-F$ units labeled 1 for intelligibility. How important is the binary pattern of an ideal mask itself? This investigation was designed to isolate the intelligibility contribution of an IBM by removing the target speech signal from all $T-F$ units.

Specifically, with linear filters, including gammatone filters (Patterson *et al.*, 1988; Wang and Brown, 2006), increasing or decreasing the SNR of a mixture while changing LC by the same amount does not alter the IBM. On the other hand, although co-reducing input SNR and LC does not change the IBM, the masked mixture becomes progressively noisier or contains less target signal. Taking this manipulation to the limit, i.e., setting both mixture SNR and LC to $-\infty$ dB, leads to an output that contains only noise with no target speech at all. This particular output corresponds to turning on or off the filtered noise according to a pattern prescribed by the IBM. Our study evaluates speech intelligibility of noise gated by the IBM obtained in this way.

[a)]Author to whom correspondence should be addressed. Electronic mail: dwang@cse.ohio-state.edu

II. METHOD

A. Stimuli

Our tests use sentences from the Dantale II data set as target speech and a speech-shaped noise as interference (Wagener *et al.*, 2003). The speech material in the Dantale II corpus consists of sentences recorded by a female Danish speaker. Each sentence has five words with a fixed grammar (name, verb, numeral, adjective and object), e.g., "Michael had five new plants" (English translation). Each position in a sentence takes a randomly chosen word from ten equally meaningful words. The speech-shaped noise included in the Dantale II corpus is produced by adding repeated utterances of each of the 250 test sentences in the corpus (see Wagener *et al.*, 2003). Both speech and noise materials were digitized at 20 kHz sampling frequency.

A speech utterance and the noise are first processed by a gammatone filterbank with varying numbers of filter channels. With 32 filters equally spaced on the equivalent rectangular bandwidth (ERB) rate scale with center frequencies distributed in the range of 2–33 ERBs (or 55–7743 Hz), the frequency response of the filterbank is nearly flat. In addition to a 32-channel gammatone filterbank, we also tested 16-, 8-, and 4-channel filterbanks. Each of the filterbanks spans the same frequency range with filters equally spaced on the ERB-rate scale, and in all cases each filter has the bandwidth of 1 ERB. With reduced channels, the frequency response of a filterbank is no longer flat and information in certain frequency bands is lost in comparison to the 32-channel filterbank case. A filter response is then windowed into time frames using 20 ms rectangular windows and a frame shift of 10 ms. This 100 Hz frame rate is commonly used in speech processing (Rabiner and Juang, 1993). The resulting $T-F$ representation has been called a cochleagram (Wang and Brown, 2006). The IBM is constructed from the cochleagrams of the target speech and the speech-shaped noise with both the mixture SNR (calculated during the period of a sentence) and LC set to 0 dB. The IBM is then applied to the noise cochleagram alone in a synthesis step to generate a waveform stimulus [see Wang and Brown (2006) for details of cochleagram analysis and synthesis]. Figure 1 illustrates the signal processing scheme using a Dantale II sentence. Take, for example, the 8-channel filterbank case. Figure 1(G) shows the IBM for this case, which is produced by comparing the 8-channel cochleagram of the Dantale II sentence and the 8-channel cochleagram of the speech-shaped noise. Applying the binary mask in Fig. 1(G) to gate the noise results in a waveform signal, which is represented in the cochleagram format in Fig. 1(H). Note that Fig. 1 represents the waveform signals from different channel numbers using the same 32-channel cochleagram representation in order to facilitate comparison. In other words, all the cochleagrams in Fig. 1 serve the purpose of signal representation and do not indicate the size of the filterbank used in IBM construction.

B. Subjects

Twelve normal-hearing, native Danish-speaking listeners participated in the experiment. All listeners had normal hearing, i.e., their hearing thresholds did not exceed 20 dB HL, and their ages ranged from 26 to 51 with the average age of 36.

C. Procedure

In each condition of the experiment, two lists, each with ten sentences, were randomly selected from the Dantale II corpus for IBM construction. Speech-shaped noise gated by the IBM was then presented to a listener. The subjects were instructed to repeat as many words as they could after listening to each stimulus corresponding to one sentence, and no feedback was provided to them regarding whether their responses were correct or not. A stimulus was presented only once. Subjects were given a training session by listening to two lists of clean (or unprocessed) sentences, which were not included in the subsequent test. Each subject test had four conditions corresponding to the filterbanks with 4, 8, 16, and 32 channels. The four test conditions plus training took less than 30 min. The presentation order of the four conditions was randomized and balanced among the 12 listeners.

Speech and noise were both set to the sound pressure level of 70 dB initially. To account for level differences caused by the use of different-sized filterbanks, stimuli were scaled by factors of two, four, and eight, for the 16-channel, the 8-channel, and the 4-channel filterbank, respectively. This level calibration resulted in stimuli with approximately the same loudness. On each trial, a stimulus was generated by the built-in sound card (SoundMAX) in a control computer (IBM ThinkCenter S50) and then presented diotically to a listener through headphones (Sennheiser HD 280 Pro) in a sound treated hearing test room.

III. RESULTS

Figure 2 shows the word recognition performance for all four conditions. The mean speech intelligibility scores for the four conditions are: 7.75%, 54.25%, 92.92%, and 97.08%, with increasing number of filter channels. The results show that nearly perfect speech recognition is obtained with 32 channels, and a high recognition rate is obtained with 16 channels. The subjects recognized more than half of the words when the number of channels was set to 8, but were unable to perform the recognition task when the number of channels was 4. A repeated measures analysis of variance (ANOVA) was conducted and the effect of number of channels was significant, $F(3,33)=179.05$, $p<0.00001$. The Tukey honest significant difference (HSD) procedure revealed that all pairwise differences among the means were significant, $p<0.001$, except for the difference between 16 and 32 channels, which was not significant. Both ANOVA and post hoc Tukey HSD tests were conducted on the rationalized arcsine-transformed percentage scores (Studebaker, 1985).

The performance variability across different listeners was small for the 32-channel and the 16-channel cases, suggesting that the acoustic information was sufficient for them to perform the recognition task. On the other hand, the individual variability for the 8-channel case was significantly larger than the 16-channel case, $F(1,11)=5.50$, $p<0.01$, sug-

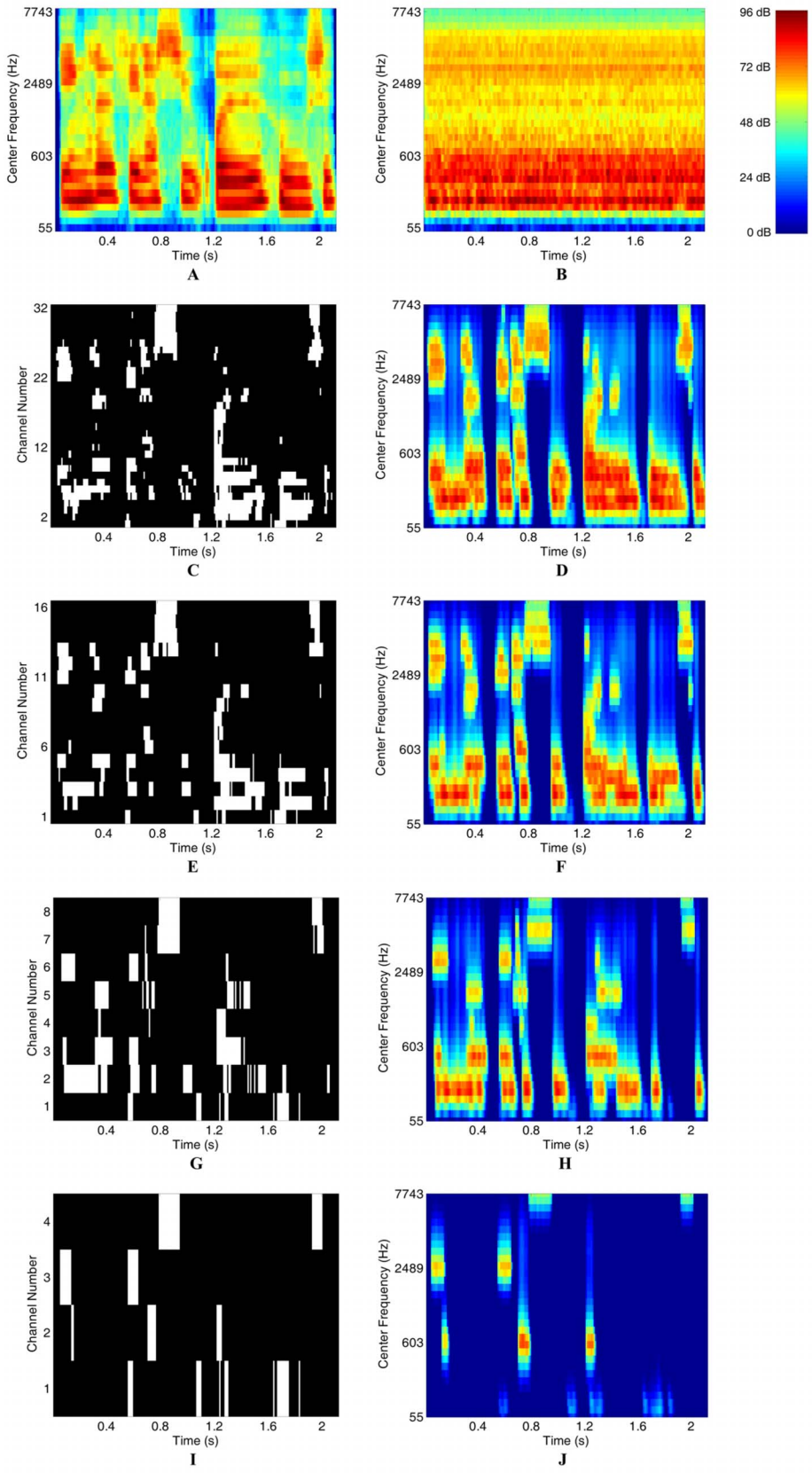

FIG. 1. (Color online) Illustration of gated noise by binary gains. (A) 32-channel cochleagram of a Dantale II sentence. (B) 32-channel cochleagram of speech-shaped noise. (C) IBM with 32 channels, where 1 is indicated by white and 0 by black. (D) 32-channel cochleagram of gated noise by the IBM in (C). (E) IBM with 16 channels. (F) 32-channel cochleagram of gated noise by the IBM in (E). (G) IBM with 8 channels. (H) 32-channel cochleagram of gated noise by the IBM in (G). (I) IBM with 4 channels. (J) 32-channel cochleagram of gated noise by the IBM in (I).

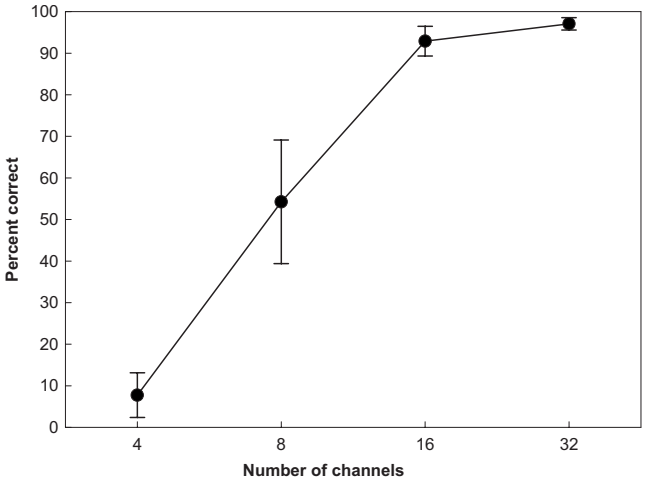

FIG. 2. Word intelligibility scores for 12 normal-hearing listeners with respect to different filterbank sizes. Dots denote the mean scores and vertical bars indicate 95% confidence intervals.

gesting that factors such as the ability and tendency to guess, concentration, and prior experience with corrupted speech, come into play.

The results in Fig. 2 clearly demonstrate that very high recognition can be obtained by turning on and off 16 bands of noise at a rate of 100 Hz following a specific pattern. The speech signal plays the sole role of determining the IBM. Such a stimulus contains little speech-specific information. The spectral shape of speech is drastically reduced to a binary variation, and so is the temporal envelope. The harmonic structure of voiced speech is absent, and the temporal fine structure (the carrier signal underlying the temporal envelope) of the stimulus reflects that of noise, not speech. Despite this dramatic reduction of speech information, listeners are capable of speech perception.

So what cues enable listeners to perceive speech from IBM-gated noise? The binary pattern encodes a general outline of spectrotemporal energy variations of speech relative to noise. Binary-gated noise crudely reflects the formant structure of speech; as shown in Fig. 1, IBM-gated noise appears to "carve out" regions of noise energy that roughly match the spectrotemporal peaks of speech. Our results indicate that such a pattern of energy variations is sufficient for recognition purposes.

IV. DISCUSSION AND CONCLUSION

Our study bears resemblance to the well-known study by Shannon *et al.* (1995) demonstrating that only a few bands of noise modulated by the corresponding speech envelopes suffice for speech intelligibility [for a much earlier study using more bands see Dudley (1939)]. There are, however, several differences between our binary-gated noise and the vocoded noise of Shannon *et al.* Perhaps the most important and obvious difference is that, within a frequency channel, noise modulation uses a binary envelope in our study and a full envelope in vocoded noise. Second, the IBM is derived by a comparison between target speech and speech-shaped noise, while temporal envelopes in vocoded noise are obtained from target speech alone. We note that many speech separation algorithms compute a binary mask by implicitly or explicitly exploiting local SNR (Divenyi, 2005; Wang and Brown, 2006), making the ideal mask amenable to computational estimation. Third, the bandwidths of noise bands in Shannon *et al.* change as the number of the bands varies in order to cover the entire frequency range of interest; in IBM-gated noise, the bandwidth of each frequency channel is fixed to 1 ERB regardless of the number of filtbank channels. It is also worth mentioning that recognizing vocoded noise takes hours of training, while no training on binary-gated noise was given in our experiment.

Like vocoded noise, the type of noise used in binary gating likely has an effect on speech intelligibility. The speech-shaped noise used in this study is a steady noise with a long-term spectrum matching that of the utterances in the Dantale II corpus, and may be particularly effective for IBM gating, although our informal listening indicates that other types of steady noise, such as pink noise, can also produce intelligible speech. Our experiment was conducted using Danish utterances. Byrne *et al.* (1994) reported that the long-term average speech spectra of a group of languages, including Danish and English, are quite similar, suggesting that, though there are likely some language effects, the main observations of our experiment may hold for English and other languages. Also, the IBM used in this study is constructed when input SNR and LC are set to be equal ($-\infty$ dB). Fixing one of them while varying the other produces different IBMs. For example, when input SNR is set to 0 dB, increasing LC results in ideal masks with fewer and fewer 1's, whereas decreasing LC leads to more and more 1's. Is equating input SNR and LC most effective for intelligibility of IBM-gated noise? Further investigation is required to address the issues regarding noise type, language, and input SNR and LC levels.

That a pattern of binary gains is apparently sufficient for human speech recognition, like previous work on vocoded noise, raises intriguing questions on the nature of human speech recognition. What speech information is truly indispensable for intelligibility? Could the IBM itself be what is being recognized? Almost perfect speech recognition from such drastically reduced speech information has broad implications for CASA, automatic speech recognition, hearing prosthesis, and coding and compression in speech communication.

ACKNOWLEDGMENTS

The authors thank Associate Editor Ken Grant, and two anonymous reviewers for their constructive criticisms and suggestions. The work was performed while D.L.W. was a visiting scholar at Oticon A/S. The authors thank M. Schlaikjer, L. Bramsløw, and M. Hartvig for their assistance in the experiments, and Y. Li for his assistance in figure preparation. D.L.W. was supported in part by AFOSR Grant No. F49620-04-01-0027 and NSF Grant No. IIS-0534707.

Allen, J. B. (**2005**). *Articulation and Intelligibility* (Morgan & Claypool, San Rafael, CA).

Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (**2006**). "Determination of the potential benefit of time-frequency gain

manipulation," Ear Hear. **27**, 480–492.
Brungart, D., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.
Byrne, D., Dillion, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Kotby, M. N., Nasser, N. H. A., El Kholy, W. A. H., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., Sirimanna, T., Tavartkiladze, G., Frolenkov, G. I., Westerman, S., and Ludvigsen, C. (**1994**). "An international comparison of long-term average speech spectra," J. Acoust. Soc. Am. **96**, 2108–2120.
Divenyi, P., ed. (**2005**). *Speech Separation by Humans and Machines* (Kluwer Academic, Norwell, MA).
Dudley, H. (**1939**). "Remaking speech," J. Acoust. Soc. Am. **11**, 169–177.
Hu, G., and Wang, D. L. (**2004**). "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Netw. **15**, 1135–1150.
Li, N., and Loizou, P. C. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.
Li, Y., and Wang, D. L. (**2008**). "On the optimality of ideal binary time-frequency masks," in *Proceedings of IEEE ICASSP*, pp. 3501–3504.
Lippmann, R. P. (**1997**). "Speech recognition by machines and humans," Speech Commun. **22**, 1–16.
Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., and Rice, P. (**1988**). "SVOS final report, part B: Implementing a gammatone filterbank," Rep. No. 2341, MRC Applied Psychology Unit.
Rabiner, L. R., and Juang, B. H. (**1993**). *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ).
Roman, N., Wang, D. L., and Brown, G. J. (**2003**). "Speech segregation based on sound localization," J. Acoust. Soc. Am. **114**, 2236–2252.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.
Wagener, K., Josvassen, J. L., and Ardenkjær, R. (**2003**). "Design, optimization and evaluation of a Danish sentence test in noise," Int. J. Audiol. **42**, 10–17.
Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell, MA), pp. 181–197.
Wang, D. L., and Brown, G. J., eds. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley, Hoboken, NJ).
Yilmaz, O., and Rickard, S. (**2004**). "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process. **52**, 1830–1847.