Effect of spectral resolution on the intelligibility of ideal binary masked speech

Ning Li and Philipos C. Loizou
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas 75083-0688
nxl051000@utdallas.edu; loizou@utdallas.edu

Abstract: Most binary-mask studies assume a fine time–frequency representation of the signal that may not be available in some applications (e.g., cochlear implants). This study assesses the effect of spectral resolution on intelligibility of ideal-binary masked speech. In Experiment 1, speech corrupted in noise at −5 to 5 dB signal-to-noise ratio (SNR) was filtered into 6–32 channels and synthesized using the ideal binary mask. Results with normal-hearing listeners indicated substantial improvements in intelligibility with 24–32 channels, particularly in −5 dB SNR. Results from Experiment 2 indicated that having access to the ideal binary mask in the F1/F2 region is sufficient for good performance.

PACS numbers: 43.72.Ar, 43.72.Ct [DO]
Date Received: December 18, 2007 **Date Accepted:** January 25, 2008

1. Introduction

The ideal binary mask (IdBM) has been set as a computational goal in computational auditory scene analysis algorithms (Wang, 2005) and has also been used extensively in "missing feature" speech recognition techniques (Cooke *et al.*, 2001). The ideal binary mask takes values of zero and one, and is constructed by comparing the local signal-to-noise ratio (SNR) in each time–frequency (T–F) bin against a preset threshold. The construction of the ideal binary mask requires knowledge of the signals (speech and interferer) prior to mixing. It is usually applied to the time–frequency representation of a mixture signal and eliminates portions of a signal (those assigned to a "zero" value) while preserving others (those assigned to a "one" value). A number of studies demonstrated high gains in speech intelligibility using the IdBM technique (Roman *et al.*, 2003; Brungart *et al.*, 2006; Li and Loizou, 2007, 2008). In the Brungart *et al.* (2006) study, for instance, performance was restored to the level attained in quiet when the IdBM technique was applied to a closed-set word test at −3 dB SNR (speech stimuli were corrupted by competing talkers).

The ideal binary mask was applied in the above-mentioned studies to the mixture signals assuming a fine time–frequency representation of the signal. The studies by Brungart *et al.* (2006) utilized a bank of 128 gammatone filters with auditory-like frequency resolution, while the study by Li and Loizou (2007) utilized a 512-point fast Fourier transform (256 channels). In applications such as hearing aids or cochlear implants, however, the time–frequency representation of the signal can be rather coarse (Loizou, 1998). In cochlear implants, for instance, speech is processed via a small number (12–22) of channels. Thus, it is not clear whether the ideal binary mask technique can bring substantial intelligibility gains, if any, when the T–F representation is poor as it is for instance in cochlear implants. Experiment 1 assesses the impact of spectral resolution on the intelligibility of IdBM speech. The spectral resolution was systematically varied by bandpass filtering speech into 6–32 channels and synthesizing it using a sinewave-excited vocoder. Since in practice algorithms that estimate the binary mask might not be accurate in all frequencies (channels), we assess in Experiment 2 the impact of frequency location of the ideal binary mask by restricting access to the ideal binary mask to a subset of channels.

2. Experiment 1: Effect of spectral resolution

2.1 Subjects and material

Fourteen normal-hearing listeners participated in this experiment. All subjects were native speakers of American English, and were paid for their participation. The speech material consisted of sentences taken from the IEEE database (IEEE, 1969). All sentences were produced by a male speaker. The sentences were recorded in a sound-proof booth (Acoustic Systems) in our lab at a 25 kHz sampling rate. Details about the recording setup and copies of the recordings are available in Loizou (2007). Two types of masker were used. The first was continuous (steady-state) noise, which had the same long-term spectrum as the test sentences in the IEEE corpus. The second masker was multitalker babble which was taken from the Auditec CD (St. Louis). The maskers were added to the target stimuli at −5, 0, and 5 dB SNR levels.

2.2 Signal processing

The stimuli were processed via an n-channel sinewave-excited vocoder (Loizou *et al.*, 1999) and synthesized with and without utilizing the ideal binary mask. In the baseline vocoder condition, signals were first processed through a preemphasis filter (2000 Hz cutoff), with a 3 dB/octave rulloff, and then bandpassed into n frequency bands ($n=6$, 12, 16, 24, and 32) using sixth-order Butterworth filters. Logarithmic filter spacing was used for $n \leqslant 16$ and mel filter spacing (linear up to 1 kHz and logarithmic thereafter) was used for higher number ($n > 16$) of channels. The envelope of the signal was extracted by full-wave rectification and low-pass filtering (second-order Butterworth) with a 400 Hz cutoff frequency. Sinusoids were generated with amplitudes equal to the rms energy of the envelopes (computed every 4 ms) and frequencies equal to the center frequencies of the bandpass filters. The sinusoids of each band were finally summed and the level of the synthesized speech segment was adjusted to have the same rms value as the original speech segment.

The stimuli were also processed and synthesized via the ideal binary mask as follows. The masker signal is first scaled (based on the rms energy of the target) to obtain the desired SNR level. The target and masker signals are then independently bandpass filtered as before into n channels (same frequency spacing), and envelopes are extracted by low-pass filtering (400 Hz cutoff) the rectified waveforms. The filtered target and masker signals are used to estimate the (true) instantaneous envelope SNR in each channel (the SNR is computed, every 4 ms, as the ratio of the rms energies of the target and masker envelope signals). If the SNR in a given channel is found to be greater than 0 dB, then the mixture envelope of that channel is retained (the 0 dB SNR threshold is adopted in this study as it is the threshold typically used for constructing ideal binary masks, Wang, 2005). If the SNR in a given channel is found to be less or equal to 0 dB, then the mixture envelope of that channel is discarded. Following the retention/discarding of the mixture envelopes in each channel, the signal is synthesized as a sum of m ($m \leqslant n$) sine waves with amplitudes set to the envelopes with positive SNR values and frequencies set to the center frequencies of the corresponding bandpass filters.

2.3 Procedure

The experiments were performed in a soundproof room (Acoustic Systems, Inc) using a PC connected to a Tucker-Davis system 3. Stimuli were played to the listeners monaurally through Sennheiser HD 250 Linear II circumaural headphones at a comfortable listening level. Prior to the test, each subject listened to vocoded speech to become familiar with the stimuli. The training session lasted for about 15–20 min. During the test, the subjects were asked to write down the words they heard. Subjects participated in a total of 60 conditions (=3 SNR levels $\times$ 2 algorithms $\times$ 2 maskers $\times$ 5 number of channels). Two different groups of subjects (seven in each group) were used due to the limited number of lists available in the IEEE corpus. The first group participated in the −5 and 0 dB conditions, and the second group participated in the 5 dB SNR conditions. Subjects were randomly assigned to the two groups. Two lists of sentences (i.e., 20 sentences) were used for each condition.[1] The sentence lists were counterbal-

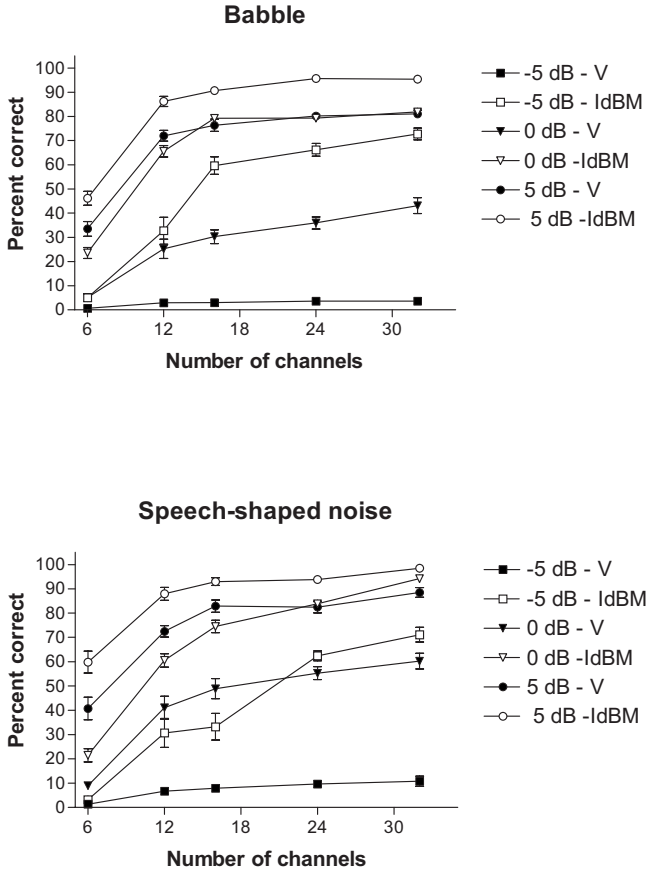

Fig. 1. Mean percent correct scores as a function of number of channels for the two types of maskers tested. The scores for the baseline vocoded stimuli (V) are shown in closed symbols and the scores for the IdBM-vocoded stimuli are shown in open symbols. Error bars indicate standard errors of the mean.

anced across subjects. Sentences were presented to the listeners in blocks, with 20 sentences/block for each condition. The different conditions were run in random order for each listener.

2.4 Results and discussion

The mean percent correct scores (all words were used in the scoring) are shown in Fig. 1 as a function of number of channels. The scores for the baseline vocoded stimuli are shown in closed symbols and the scores for the IdBM-vocoded stimuli are shown in open symbols. For the stimuli corrupted by multitalker babble, three-way analysis of variance (ANOVA) (with repeated measures) indicated significant effect [$F(4,24)=337.9$, $p<0.005$] of spectral resolution (number of channels), significant effect [$F(1,6)=476.0$, $p<0.005$] of processing (vocoding and IdBM vocoding), significant effect [$F(2,12)=720.5$, $p<0.005$] of SNR level (−5, 0, and 5 dB) and significant ($p<0.005$) interactions among all factors. Similar ANOVA was applied to the speech-shaped noise conditions. Analysis indicated significant effect [$F(4,24)=268.5$, $p<0.005$] of spectral resolution (number of channels), significant effect [$F(1,6)=248.8$, $p<0.005$] of processing (vocoding and IdBM vocoding), significant effect [$F(2,12)=359.4$, $p<0.005$] of SNR level (−5, 0, and 5 dB), and significant ($p<0.005$) interactions among all factors.

As expected, performance improved as the number of channels increased for both baseline and IdBM vocoded stimuli. Performance reached a plateau, however, in most conditions depending on the masker and SNR level used. Post-hoc tests (according to Scheffé) were run to find the number of channels needed to reach asymptotic performance. For the IdBM stimuli processed in babble, performance reached an asymptote with 16 channels of stimulation in the −5 and 0 dB SNR conditions. In the 5 dB SNR condition, performance obtained with 12 channels did not differ significantly ($p > 0.05$) from performance obtained with 16 channels. For the IdBM stimuli processed in speech-shaped noise, performance asymptoted at 24 channels of stimulation in −5 dB SNR, and at 12 channels of stimulation in 5 dB SNR. There was no asymptote in the 0 dB SNR condition.

The above-presented analysis clearly indicates that spectral resolution has a significant impact on the intelligibility of IdBM stimuli. The performance with the IdBM stimuli did not reach the same level (90%–100% correct) as attained by Brungart *et al.* (2006) with 128 channels in low (−3 dB) SNR conditions. Nevertheless, the improvement in performance obtained with the IdBM stimuli in the present study is quite substantial, particularly at low SNR levels (−5 and 0 dB). In the −5 dB SNR babble condition, for instance, performance improved by roughly 60 percentage points with 24–32 channels of stimulation.

3. Experiment 2: Effect of frequency location of binary mask

In the previous experiment we assumed that we had access to the ideal binary mask in all channels. In practice, the binary mask (or equivalently the envelope SNR) needs to be estimated from the noisy observations and the SNR estimation may not be accurate in all channels. Hence, in the present experiment we assess the impact of frequency location of the binary mask on speech intelligibility. This is done by assuming access to the ideal binary mask for only a subset of the channels and leaving the remaining channels unaltered.

3.1 Subjects and material

A different group of six normal-hearing listeners participated in this experiment. All subjects were native speakers of American English, and were paid for their participation. The same speech material from the IEEE database was used for the target stimuli, and the same types of maskers were used as in Experiment 1. The maskers were added to the target stimuli at −5, 0, and 5 dB SNR levels.

3.2 Signal processing

The stimuli were processed as before into 32 channels via the ideal binary mask technique, except for the following difference in implementation. Access to the ideal binary mask was restricted only to channels falling within a limited frequency region spanning 0–f Hz, where f=560, 1000, 1630, 2720, and 5500 Hz. Hence, channels falling within the 0–f Hz region were synthesized using the IdBM technique (same as in Experiment 1) and channels with upper cutoff frequencies greater than f Hz were synthesized via the baseline vocoding strategy. Note that the f=5500 Hz condition corresponds to the condition in which all channels had access to the ideal binary mask (same as in Experiment 1), and is included here for comparative purposes. The above-mentioned cutoff frequencies were chosen to assess the importance of having access to F1 information ($f < 1000$ Hz) alone or F1/F2 information ($f < 2720$ Hz) alone.

3.3 Procedure

The same experimental setup was used as in Experiment 1. Subjects participated in a total of 36 conditions (=3 SNR levels × 6 cutoff frequencies × 2 maskers) including the baseline vocoder and the IdBM-vocoded conditions. Two lists of sentences (i.e., 20 sentences) were used per condition, and none of the lists were repeated across conditions. The sentence lists were counterbalanced across subjects. Sentences were presented to the listeners in blocks, with 20 sentences/block in each condition. The different conditions were run in random order for each listener.

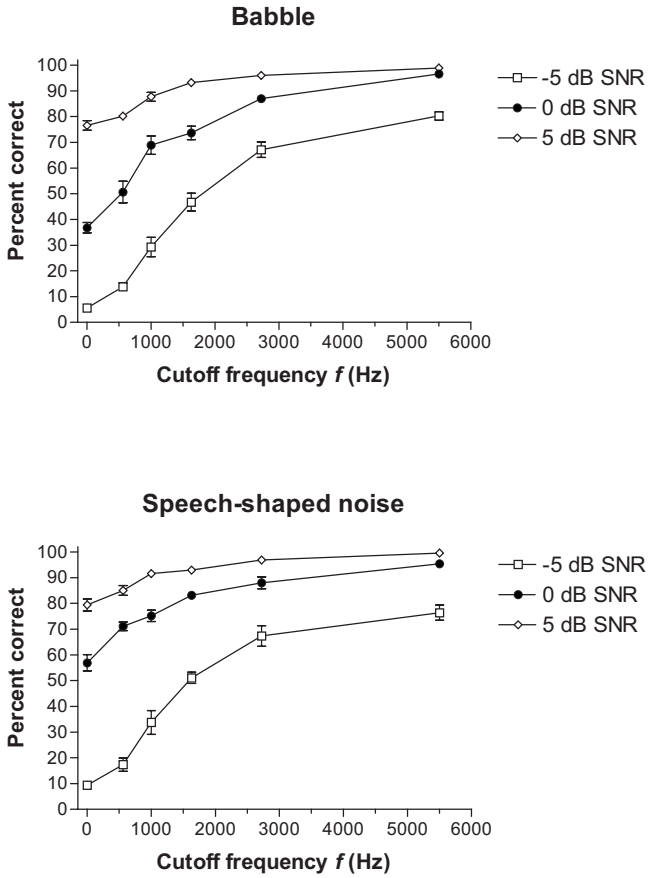

Fig. 2. Mean percent correct scores as a function of the cutoff frequency f (i.e., frequency location of the ideal binary mask) for the two types of maskers tested. The $f=0$ Hz condition corresponds to the baseline vocoded stimuli which made no use of the ideal binary mask. All stimuli were processed via 32 channels. Error bars indicate standard errors of the mean.

3.4 Results and discussion

The mean percent correct scores are shown in Fig. 2 as a function of the cutoff frequency f (Hz), i.e., the frequency location of the ideal binary mask. The $f=0$ Hz condition corresponds to the baseline vocoded stimuli (no access to the ideal binary mask) and the $f=5500$ Hz condition corresponds to the situation in which all 32 channels had access to the ideal binary mask (as in Experiment 1). For the stimuli corrupted by multitalker babble, two-way ANOVA (with repeated measures) indicated significant effect [$F(5,25)=269.5$, $p<0.005$] of the cutoff frequency, significant effect [$F(2,10)=862.3$, $p<0.005$] of SNR level, and significant interaction [$F(10,50)=35.7$, $p<0.005$]. For the stimuli corrupted by speech-shaped noise, two-way ANOVA (with repeated measures) indicated significant effect [$F(5,25)=266.3$, $p<0.005$] of the cutoff frequency, significant effect [$F(2,10)=492.4$, $p<0.005$] of SNR level, and significant interaction [$F(10,50)=32.9$, $p<0.005$].

As shown in Fig. 2, performance improved monotonically as the cutoff frequency f increased (i.e., more channels were included with access to the ideal binary mask). The above-presented statistical analysis yielded an interaction between SNR level and cutoff frequency, and that interaction stems from the fact that the rate of improvement in performance differed for the three SNR levels. A steep rate of improvement in intelligibility was observed for low SNR

levels (−5 dB) and a relatively shallow rate of improvement was observed for higher SNR levels (5 dB) as f increased (i.e., as more channels with ideal binary mask were added). Post-hoc tests, according to Scheffé, indicated that performance asymptoted at f=2500 Hz in the −5 and 0 dB SNR conditions for both maskers. Performance asymptoted at f=1500 Hz in the 5 dB SNR condition for both maskers.

The outcome from the present experiment suggests that it is important for algorithms that estimate the binary mask to be accurate in the low frequency region, and in particular the F1/F2 region. The intelligibility tests in Experiment 2 suggest that having access to the ideal binary mask in the low frequencies is sufficient for good performance. It is speculated that access to a better SNR in the low-frequency region makes it easier for listeners to segregate the target in complex listening situations via a glimpsing mechanism. Evidence of the advantage introduced by glimpsing the low-frequency region was provided in the study by Li and Loizou (2007) and Anzalone *et al.* (2006). Significant reductions in speech reception threshold were obtained in the study by Anzalone *et al.* (2006) by both normal-hearing and hearing-impaired listeners when the ideal speech detector was applied only to the lower frequencies (70–1500 Hz).

4. Conclusions

The present study extended previous findings on the intelligibility of ideal binary masked speech (Brungart *et al.*, 2006; Li and Loizou, 2008). The present results indicate that the use of ideal binary masks can bring substantial gains in speech intelligibility, particularly at low SNR levels (−5 and 0 dB), even when the spectral resolution is relatively low (16–24 channels). Access to the ideal binary mask in the low frequencies, particularly in the F1/F2 region, was found to be sufficient for good performance.

Acknowledgments

This research was supported by Grant No. R01 DC007527 from the National Institute of Deafness and other Communication Disorders, NIH.

References and links

[1]Due to the limited number of IEEE sentence lists, eight lists had to be reused in four conditions. These eight lists, however, were chosen from the −5dB SNR babble conditions, in which subjects performed very poorly (<5% correct). This was done to avoid learning effects.

Anzalone, M., Calandruccio, L., Doherty, K., and Carney, L. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation." Ear Hear. **27**, 480–492.
Brungart, D., Chang, P., Simpson, B., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation." J. Acoust. Soc. Am. **120**, 4007–4018.
Cooke, M. P., Green, P. D., Josifovski, L., and Vizinho, A. (**2001**). "Robust automatic speech recognition with missing and uncertain acoustic data." Speech Commun. **34**, 267–285.
IEEE (**1969**). "IEEE recommended practice for speech quality measurements." IEEE Trans. Audio Electroacoust. **17**, 225–246.
Li, N., and Loizou, P. (**2007**). "Factors influencing glimpsing of speech in noise." J. Acoust. Soc. Am. **122**, 1165–1172.
Li, N., and Loizou, P. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction." J. Acoust. Soc. Am., **123**(3) (to be published).
Loizou, P. (**1998**). "Mimicking the human ear: An overview of signal processing techniques for converting sound to electrical signals in cochlear implants." IEEE Signal Process. Mag. **15**, 101–130.
Loizou, P. (**2007**). *Speech Enhancement: Theory and Practice*, (CRC Press, Boca Raton, FL).
Loizou, P., Dorman, M., and Tu, Z. (**1999**). "On the number of channels needed to understand speech." J. Acoust. Soc. Am. **106**, 2097–2103.
Roman, N., Wang, D., and Brown, G. (**2003**). "Speech segregation based on sound localization." J. Acoust. Soc. Am. **114**, 2236–2252.
Wang, D. (**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, Dordrecht), pp. 181–197.