Ulrik Kjems[1], Michael S. Pedersen[1], Jesper B. Boldt[1], Thomas Lunner[2], DeLiang Wang[3]
[1]Oticon, Denmark, [2]Research Center Eriksholm, Denmark and Linköping Univ, Sweden, [3]Ohio State University, Ohio, USA

Oticon A/S, Kongebakken 9, Smørum DK-2765, Denmark, Email: uk@oticon.dk

Speech Intelligibility from Ideal Time-Frequency Gain Manipulations

Motivation

Understanding the mechanisms why ideal time-frequency gain manipulations increase intelligibility.

Ideal time-frequency gain manipulations

$$IBM(t,f) = \begin{cases} 1 & \text{if } S(t,f) - M(t,f) > LC \quad [dB] \\ 0 & \text{otherwise.} \end{cases}$$

Eq. 1: Definition of ideal binary gain

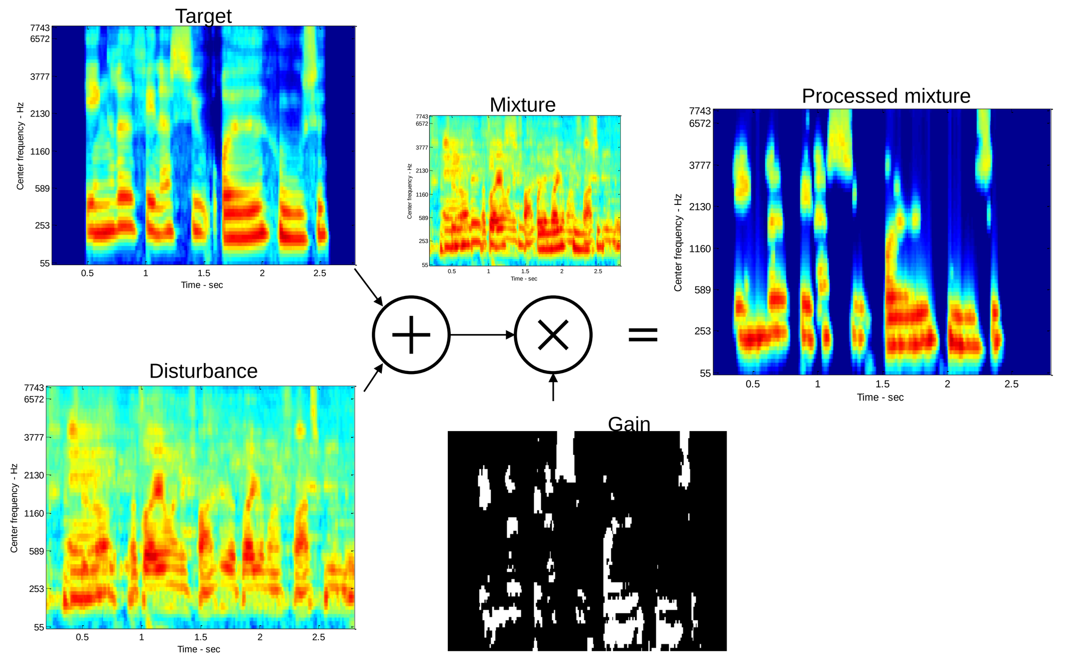


Fig. 1: Illustration of application of ideal binary gain pattern

RC = Relative Criterion

IBM is unchanged if *LC* and mixture *SNR* covary.

Therefore relative criterion (***RC***) is introduced as difference between *local* SNR and *global* mixture SNR:

$$RC = LC - SNR$$

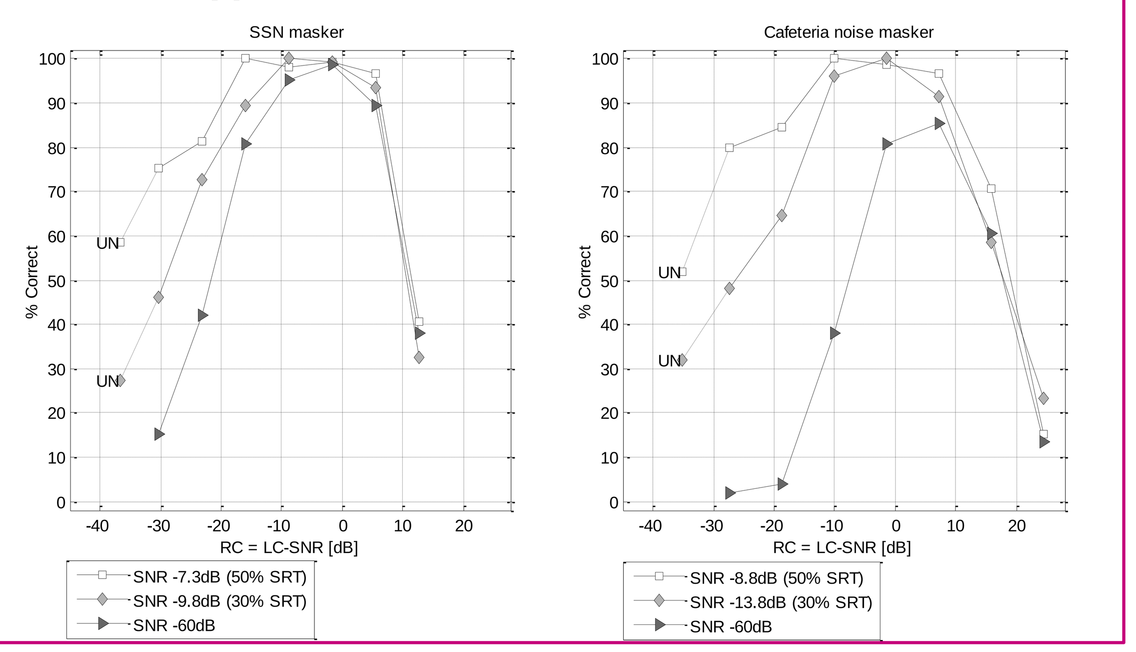
In a listening experiment, gain patterns are fixated (i.e. hold RC constant) while measuring intelligibility for varying global mixture SNR values.

In this manner effect of gain pattern is studied.

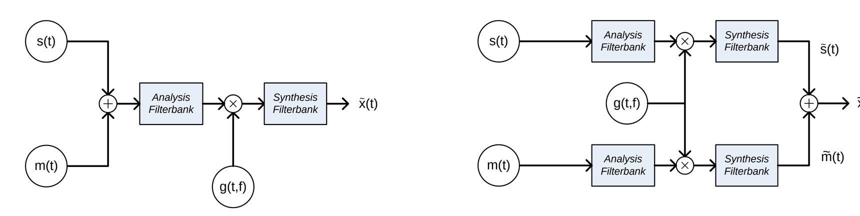
Experimental Setup

- Listening experiment with headphones
- 15 normal hearing subjects
- Dantale II sentences (5 word sentences from vocab of 10 at each place)
- Four noise types (SSN, cafeteria noise, car noise and bottling hall noise)
- IBM processed mixtures with three mixture SNR settings × 8 RC values

More details in [5].



Two assumed mechanisms for conveying intelligibility



Processed output is sum of processed target and processed disturbance

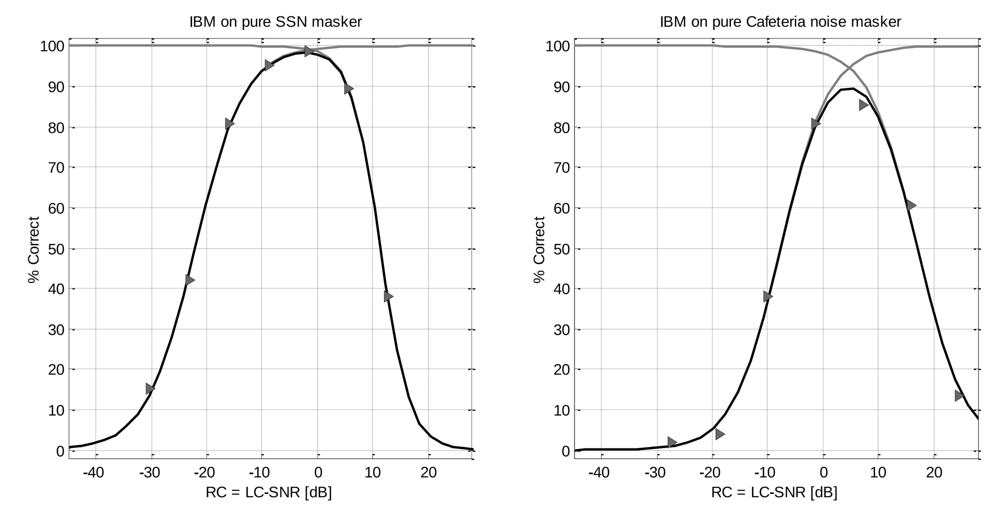
$$\tilde{x}(t) = \tilde{s}(t) + \tilde{m}(t)$$

Assumed equivalent to two processes for conveying intelligibility:

$\tilde{m}(t)$ — Processed noise

$\check{s}(t) = \tilde{s}(t) + m(t)$ — Processed clean speech plus *unprocessed* noise

Intelligibility of processed noise $\tilde{m}(t)$



$$\mathcal{I}_{\tilde{m}}(RC) = L_{vocoder}(RC) \cdot L_{sparsity}(RC)$$

	SSN		Cafeteria	
Masker	***s***	***r***	**s**	***r***
$L_{sparsity}$	-0.094dB^{-1}	11.4dB	-0.058dB^{-1}	17.1dB
$L_{vocoder}$	0.056 dB^{-1}	-22.1dB	0.059 dB^{-1}	-7.6dB

Table 1: Logistic parameters for $\tilde{m}(t)$

Masker	L_{50}	s_{50}
SSN	-7.3 dB	13.2 %/dB
Cafeteria	-8.8 dB	6.8 %/dB

Table 2: Logistic parameters for psychometric function.

Intelligibility of processed clean speech plus unprocessed noise $\check{s}(t)$

$$\mathcal{I}_{\check{s}}(SNR, RC) = \mathcal{I}(SNR) \cdot L_{sparsity}(RC)$$

$$\mathcal{I}(SNR) = \{1 + \exp[4 s_{50}(L_{50} - SNR)]\}^{-1}$$

$$L_{sparsity}(RC) = \{1 + \exp[4 s_{sparsity}(r_{sparsity} - RC)]\}^{-1}$$

Two independent channels of information, a missed word means error in both:

$$1 - \mathcal{I}_{\tilde{x}}(SNR, RC) = (1 - \mathcal{I}_{\check{s}}(SNR, RC))(1 - \mathcal{I}_{\tilde{m}}(RC))$$

Final Model

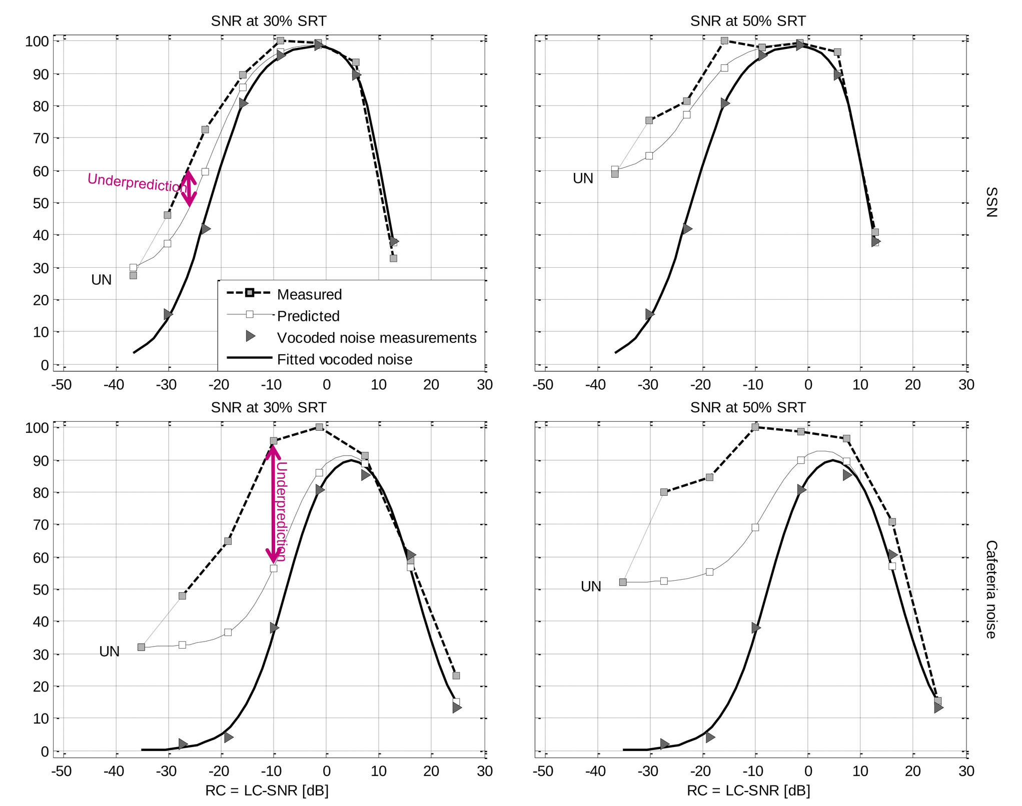
$$\mathcal{I}_{\tilde{x}}(SNR, RC) =$$

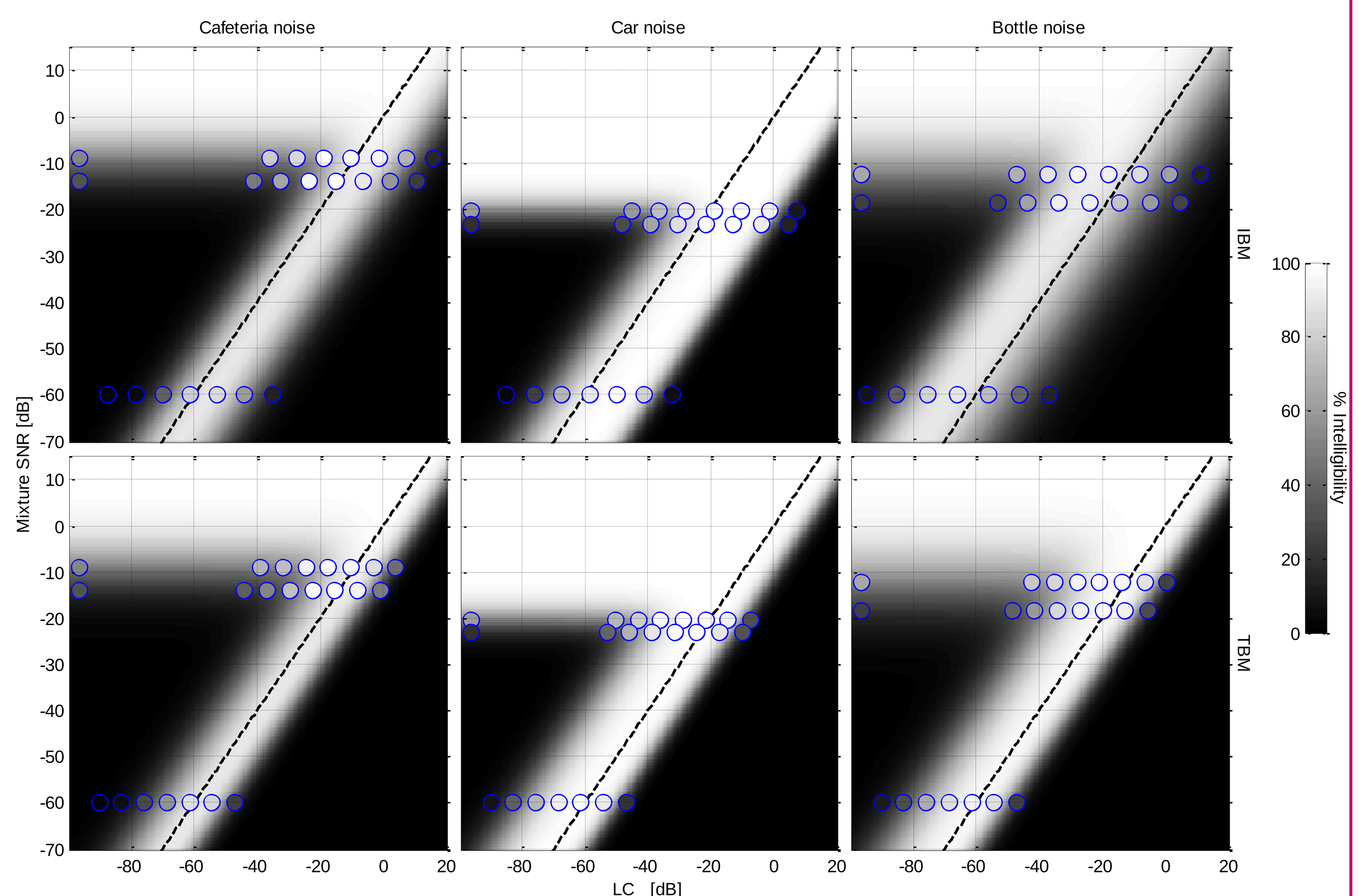
$$[\mathcal{I}(SNR) + L_{vocoder}(RC) - \mathcal{I}(SNR) \cdot L_{vocoder}(RC) \cdot L_{sparsity}(RC)] \cdot L_{sparsity}(RC)$$

This model predicts intelligibility of IBM processed mixtures from knowledge of intelligibility of vocoded noise, and of the psychometric curve for additive noise.

Model Predictions

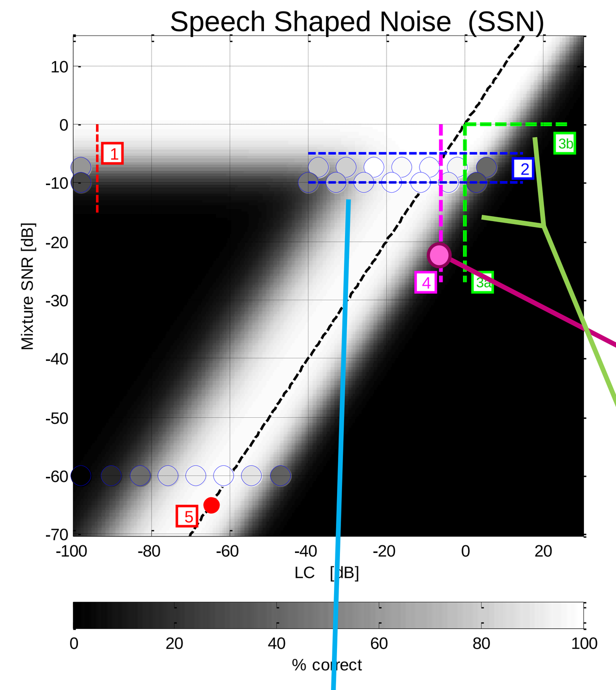
Comparison of measured and predicted intelligibilities





- Performance for sparse masks is accurately predicted by model
- Underprediction for dense masks (low RC values) suggesting a reduction of informational masking by IBM processing

Model Predictions in (LC,SNR) plane



1. Standard psychometric curve for additive noise
2. Li&Loizou 2008,JASA 123
3. Brungart et al. 2008, JASA 120
4. Wang et al. 2009, JASA 125
5. Wang et al. 2008, JASA 124

Noise type	Measured	Predicted
SSN	-17.1dB	-17.4dB
Café noise	-23.2dB	-23.1dB

Table 3: Measure vs predicted SRTs

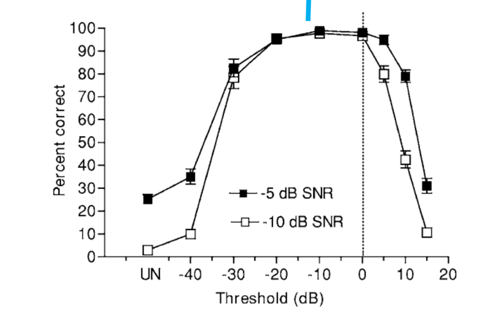
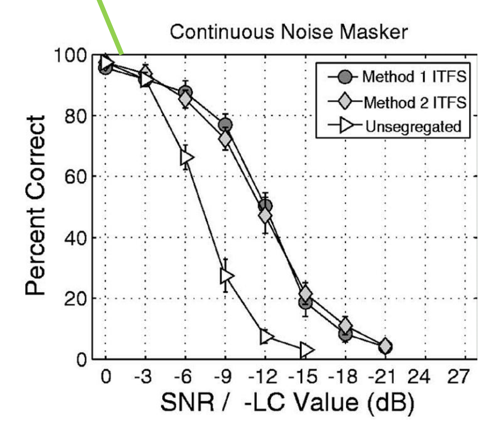


FIG. 2. Performance (percent of words identified correctly) as a function of SNR threshold (dB) for two input global SNR levels. The masker was 20-talker babble. Performance obtained with unprocessed mixtures is indicated as UN. Error bars indicate standard errors of the mean.

N. Li and P. C. Loizou: Perception of binary-masked speech 1675



Predicting SRT of IBM processed mixtures

The right-hand sloping side of "figure 7" is determined by the $r_{sparsity}$ parameter, setting $L_{sparsity}(RC)=0.5$ yields $RC = r_{sparsity}$, or

$$SRT_{IBM} \cong LC - r_{sparsity}$$

The prediction is shown in Table 3 above match the experimental data very well indeed.

Conclusions

The proposed model gives a qualitative description of where in the (LC, SNR)-space the benefits of ideal gain manipulations occur.

The model makes predictions based on recognition scores of vocoded noise, and knowledge of the psychometric curve for additive noise.

The model predicts that the optimal LC value for ideal gain manipulations depends on the mixture SNR, so that

$$LC_{opt} = SNR + RC_{opt}$$

The model does underpredict performance in some regions (with SNR near SRT with lower LC values than the above optimal value), indicating that there is more benefit than predicted by this model (i.e. more than can be explained by vocoding). This additional benefit could be explained by release of informational masking by IBM processing.

Finally, the model shows good qualitative agreement with previously published experimental data, and accurately predicts SRTs in a previous IBM experiment.

References

[1] Li, N., and Loizou, P. C., " Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. 123, 1673-1682, 2008.
[2] Brungart, D., Chang, P. S., Simpson, B. D., and Wang, D. L, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. 120, 4007-4018, 2006.
[3] Wang, D. L., Kjems, U, Pedersen, M. S., Boldt, J. B., and Lunner, T., "Speech Perception of Noise with Binary Gains". J. Acoust. Soc. Am. 124, 2303-2307, 2008.
[4] Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M., "Speech recognition with primarily temporal cues," Science 270, 303-304, 1995.
[5] Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. L., "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. 126, 1415-1426, 2009.
[6] Allen, J. B., "The Articulation Index is a Shannon channel capacity," Auditory Signal Processing, Springer New York, 313-319, 2006.
[7] Wang, D. L. U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner. "Speech Intelligibility in Background Noise with Ideal Binary Time-frequency Masking". J. Acoust. Soc. Am. 125, 2336-2347, 2009.
[8] U. Kjems, M. S. Pedersen, J. B. Boldt,T. Lunner and Wang, D. L. "Speech Intelligibility of Ideal Binary Masked Mixtures". EUSIPCO 2010.

PEOPLE FIRST