Energy reallocation strategies for speech enhancement in known noise conditions

Yan Tang[1], *Martin Cooke*[1, 2]

[1]Language and Speech Laboratory, Faculty of Letters, Universidad del Pais Vasco, Spain
[2]Ikerbasque (Basque Science Foundation)

y.tang@laslab.org, m.cooke@ikerbasque.org

Abstract

Speech output, whether live, recorded or synthetic, is often employed in difficult listening conditions. Context-sensitive speech modifications aim to promote intelligibility while maintaining quality and listener comfort. The current study used objective measures of intelligibility and quality to compare five energy reallocation strategies operating under equal energy and preserved duration constraints. Results in both stationary and highly-nonstationary backgrounds suggest that time-varying modifications lead to large increases in objective intelligibility, but that speech quality is best preserved by time-invariant modifications. Selective amplification of time-frequency regions with low *a priori* SNR produced the highest objective intelligibility without severe disruption to quality.

Index Terms: energy reallocation, speech intelligibility, glimpsing, SII, PESQ

1. Introduction

Speech output technology is frequently encountered in less-than-ideal listening conditions (e.g. public address systems, in-vehicle talking GPS). Humans appear to use their shared experience of ongoing listening conditions to modify speech output to meet the needs of listeners (e.g. [1,2]) but speech generation technology is currently context-unaware, and there is no guarantee that messages conveyed by synthetic, recorded or live speech are understood.

Several recent approaches have been proposed to enhance speech intelligibility under adverse conditions [3-6]. A dynamic range controller was introduced in [3] to increase the perceived loudness while maintaining the original intensity range. An alternative approach developed in [4] aimed to improve intelligibility by amplifying transient components extracted using a set of time-varying filters whose center frequencies and bandwidths were controlled to identify the strongest formants. Similarly, formant-enhancement was employed in [5]. Speech intelligibility was boosted in [6] by raising the average speech spectrum over the average noise spectrum using frequency-dependent and -independent signal-to-noise ratio (SNR) recovery.

Increasing SNR via amplification is clearly beneficial [4-6] but the use of excessive output levels leads to discomfort and stress [7] and may damage the hearing of those exposed for sustained periods, and can cause irreversible equipment damage [8]. Boosting signal energy uniformly across frequency and time is also an inefficient way to increase intelligibility, since the degree of amplification required to overcome energetic masking in any time-frequency region is a function of the local SNR. Indeed, reductions in speech level may be appropriate in regions of high SNR, while for those parts where the SNR is extremely low, no global increase in speech level may be feasible for unmasking. Given the redundancy of the speech signal in the time-frequency domain (e.g. [9]), it is not necessary to ensure that speech energy survives masking everywhere, suggesting the use of modification strategies that selectively amplify or attenuate regions of the speech signal. Here we explore a class of speech modification approaches which *reallocate* speech energy in time and frequency under the constraint that the overall energy is unchanged (note that this approach does not guarantee preservation of loudness). Five strategies which differ in the use of global and local information are compared using objective intelligibility and quality metrics.

2. Speech energy reallocation

The strategies explored here assume that the speech and noise signals are known (or can be estimated), and that only the speech signal can be changed. They consist of two steps: modification of speech energy, followed by speech energy renormalisation. Together, these steps achieve a reallocation of speech energy across time and frequency. In addition to the constant energy constraint, in the current work we make the assumption that the duration of the modified speech is unchanged from the original.

Energy modification is equivalent to applying a spectro-temporal weighting k_M to the energy distribution of the original signal *S,* resulting in a modified energy distribution *S'*

$$\sum_{t=1}^{T}\sum_{f=1}^{F} S'(f,t) = \sum_{t=1}^{T}\sum_{f=1}^{F} k_M(f,t)\cdot S(f,t) \qquad (1)$$

where T and F are the number of frames and frequency channels and k_M is a positive weight matrix. The energy renormalisation step corresponds to

$$\sum_{i=1}^{n} s'(i)^2 = \sum_{i=1}^{n} s(i)^2 \qquad (2)$$

where *s'* and *s* denote the modified and original speech waveforms and *n* is the number of samples.

In addition to a no-modification baseline (M0), we investigated 5 modification strategies (M1-M5). The first 3 were based on equalising local SNRs to a fixed global SNR, and differed in whether the modification was applied at the level of time frames (M1), frequency channels (M2) or time-frequency cells (M3). Two further strategies applied energy modification to a subset of frequency channels (M4), or made changes based on the local SNR (M5). In the following, SNR_{global} represents the global SNR, while *SNR(f)* and *SNR(t)* denote the SNR in a given frequency band or time frame, while *SNR(f,t)* signifies the local SNR in a time-frequency cell. Time and frequency channel indices are represented throughout by *t* and *f* with ranges 1..*T* and 1..*F*, and primes are used to distinguish SNRs of modified and original signals .

M0: ***No modification.*** Here, the signal is not changed:

$$k_{M0}(f,t) = 1, \forall f,t \qquad (3)$$

M1: ***Constant SNR per time frame.*** This approach applies a time-varying gain to the speech signal which makes the framewise SNR constant across the signal. The effect is to reallocate energy across time to those segments with adverse

SNR from those with more advantageous SNR. Formally, *SNR'(t)* is adjusted to be equal to SNR_{global} such that all k_{M1} *(f, t)* in a given time frame are identical:

$$k_{M1}(1,t) = k_{M1}(2,t) = \cdots k_{M1}(F,t) = K(t), \forall t \quad (4)$$

Window size is a key parameter. Adjustment of the SNR in short windows (e.g. ~10 ms) may improve intelligibility more than for longer ones (~100 ms) at the expense of speech quality. Using objective intelligibility and quality measures (sec. 3), we found 50 ms Hamming windows to provide a reasonable tradeoff between quality and intelligibility. This value was used for strategies M1, M3 and M5.

M2: ***Constant SNR per frequency channel.*** This applies a time-invariant spectral tilt to equalise the SNR in each frequency region. *SNR'(f)* is adjusted to SNR_{global} such that the modification is constant across time:

$$k_{M2}(f,1) = k_{M2}(f,2) = \cdots k_{M2}(f,T) = K(f), \forall f \quad (5)$$

To reflect human frequency resolution, speech was processed by a gammatone filterbank [10] with *F* channels in the range 100-7500 Hz (was applied to M2-M5). Previous applications of the gammatone filterbank have used 30-100 channels. Here, we measured the objective speech quality (sec. 3) of the reconstructed signal as a function of F in the range 32-100 channels. Quality asymptoted at F=55 channels, i.e. 1.3 filters per ERB (a value compatible with [11]). This value was chosen for strategies M2-M5. For signal reconstruction, filter outputs were reversed, refiltered and reversed again to remove phase distortions prior to summing across filters.

M3: ***Constant SNR per time-frequency cell.*** A generalisation of M1 and M2 is to modify the local SNR so that each time-frequency region of speech is equated for SNR:

$$SNR'(f,t) = SNR_{global}, \forall f, t \quad (6)$$

M4: ***Boosting of selected frequency channels.*** Frequency regions differ in their importance for speech perception, a fact reflected in the weightings for objective intelligibility measures such as the speech intelligibility index (SII) [12]. M4 operates by increasing speech level in selected frequency regions. A general version of this strategy would allow for the choice of an arbitrary subset of channels with a different gain factor for each channel, but to reduce the computational complexity of parameter optimisation, we used a fixed increase in level and a single contiguous set of channels anchored at the low or high frequency end of the filterbank:

$$SNR'(f) = SNR(f) + E_{boost}, \forall f \in C_i \quad (7)$$

where E_{boost} is the amount (in dB) by which each channel in the set of channel indices C_i is amplified. E_{boost} and C_i were optimised independently for maximal objective intelligibility based on a glimpsing metric [13] (sec. 3), resulting in the values E_{boost} = 20 dB and C_i = {33, 34, 35…55}, held constant for all utterances. This range of channels corresponds to centre frequencies from 1800-7500 Hz, a finding broadly consistent with the mid-frequency boost seen in clear speech [14] where the range 1000-3000 Hz typically contains more energy.

M5: ***Local SNR-specific boosting within selected channels.*** This strategy applies a refinement to those frequency channels selected in M4 based on the idea that energy reallocation to regions which already possess a high local SNR is unlikely to improve intelligibility and, under a constant energy constraint, is wasteful. Similarly, regions with a very low local SNR may require too great an injection of energy to make a significant contribution to intelligibility. Objective intelligibility metrics such as SII use the range of band-based SNRs of 0-30 dB to weight the contribution of frequency bands to overall intelligibility, with zero weighting for SNRs less than 0 and unity for 30 dB or greater. M5 performs selective energy boosting based on local SNR. Only those frames whose local SNR falls into a specified set of ranges *range* are boosted:

$$SNR'(f,t) = SNR(f,t) + E_{boost}, \forall SNR(f,t) \in range \quad (8)$$

Additionally, based on SII, we limit the maximum local SNR to 30 dB to reallocate 'excess' energy elsewhere:

$$SNR'(f,t) = \min(30, SNR(f,t)) \quad (9)$$

For the purposes of optimisation, the local SNR range was quantised into 8 regions in dB:

$$range = \{< 0, 0\text{-}5, 5\text{-}10, 10\text{-}15, 15\text{-}20, 20\text{-}25, 25\text{-}30, > 30\} \quad (10)$$

Using the MATLAB genetic algorithm optimisation toolbox, objective intelligibility and quality were jointly optimised, resulting in the subset to boost $range_{=}$ {< 0 d B, 0-5 dB}. This finding suggests that under a constant energy constraint, boosting those regions which make no contribution or only a marginal contribution to intelligibility is the best strategy rather than amplifying portions of the signal whose local SNR is low to moderate.

Figure 1 illustrates the effect of strategies M0-M5 for speech in the presence of a competing talker. Regions in red and blue show boosted and attenuated portions of the signal respectively, with color saturation denoting the degree of energy change. The patterns of energy shift are clearly related to the temporal, spectral or spectro-temporal nature of the strategy in M1-M3. The use of a constant energy boost in mid-to-high frequencies is evident in M4, while these channels show more variable boosting in M5 where a local SNR criterion determines whether amplification is applied.

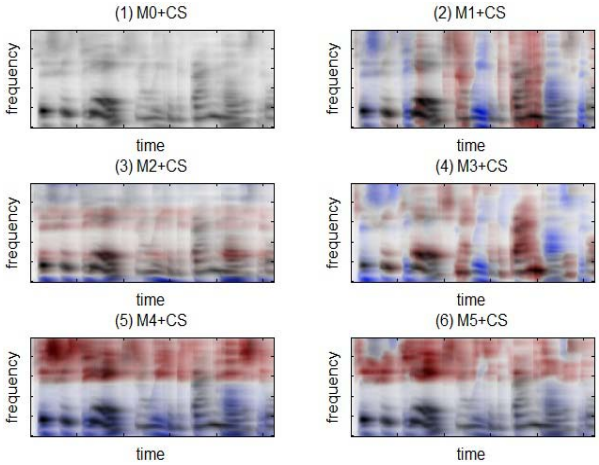

Figure 1: *Energy reallocation induced by six manipulations in the presence of a competing speaker (CS) at* SNR_{global} *= - 5dB.*

3. Intelligibility and quality metrics

The current study employed objective metrics of intelligibility and speech quality to estimate the effect of speech modification on listeners. Objective measures enable rapid feedback on a range of modification techniques and allow parameter optimisation, and can be used to select candidate approaches for subsequent formal listening tests. We used two objective intelligibility measures, SII [12] and glimpse proportion [13], and one quality metric, Perceptual Evaluation of Speech Quality (PESQ) [15, 16].

SII uses a weighted sum of channel intelligibilities, each derived from a function of the SNR in that channel. SII provides good estimates for quasi-stationary sources, but since we were also interested in the effect of common nonstationary backgrounds such as competing speech from one or more talkers, we in addition used a recent metric based on glimpses – spectro-temporal regions where the speech is more energetic than the masker – which also gives a good match to subjective intelligibility for non-stationary sources [13, 17].

The PESQ score is based on the difference between loudness spectra for the original and modified speech signals. These signals are first normalized to a standard listening level and then filtered with a frequency response similar to a standard telephone handset. To obtain loudness spectra, a time alignment is applied to the signals to correct time delays before processing via a Bark spectral distortion (BSD)-like auditory transform. PESQ values fall into the range 0.5-4.5, with higher scores corresponding to better speech quality. Subjective listening tests in [15] demonstrated a high correlation with PESQ values ($\rho = 0.92$).

4. Results

The effect of the modification strategies described in section 2 was evaluated using 120 sentences from the SCRIBE corpus [18] (60 male, 60 female) of average length 2.68s, added to competing speech (CS), 8-talker babble (BAB) or speech-shaped noise (SSN) at SNRs of -10, -5, 0, 5 and 10 dB.

Figure 2 presents glimpse proportion, SII and PESQ measures as averages over the corpus in each test condition. With the exception of M3, all methods increased intelligibility under all experimental conditions, with M5 being particularly effective. For 4 of the 5 approaches, glimpse proportion and SII produced a similar ranking of methods. M3, which maintains a constant local SNR, led to a *decrease* in model intelligibility based on glimpses at -5 and -10 dB but provides a very large increase for SNRs >= 0 dB. The difference between the SII and glimpse proportion measures stems from the fact that the glimpsing measure is highly sensitive to the local SNR while SII employs the average SNR in a frequency channel. Clearly, attempting to maintain a constant SNR in each time-frequency region is very effective as long as the SNR is positive but is a poor strategy otherwise.

Benefits of speech modification are seen across the range of SNRs. For the fluctuating maskers, modelled intelligibility increases tended to be larger for more adverse noise levels while the reverse was true for the stationary masker. This may be due to the greater availability of local opportunities to increase the audibility of speech in the presence of a time-varying background. Strategies were ranked similarly for the three backgrounds. While SII suggests similar intelligibility for the three maskers, glimpse proportions are highest for competing speech and lowest for stationary noise which correlates well with listeners [19] and highlights the lack of applicability of SII for fluctuating sources.

Different modification strategies led to a wide range of objective speech quality, with PESQ values ranging from less than 1 to levels close to the original speech (4.5). Modifications which produce temporal fluctuations (M1, M3 and M5) have the greatest negative impact on quality, and the technique which maintains a constant local SNR is particularly poor. PESQ scores favour those approaches (M2 and M4) which modify only the spectral balance of the signal. The ranking of PESQ scores for the different strategies is similar across noise backgrounds but deterioration is worst for the competing speech masker, presumably because due to increased opportunities for local energy modification and hence lower PESQ scores presented by a fluctuating masker.

PESQ scores show no effect of SNR apart from the M5 strategy, where quality decreased with noise level. Here, whether energy is boosted or not is based on the local SNR. For low global SNRs, local SNRs are in the range where boosting is applied most of the time (eqn. 8) and since boosting is by a constant dB amount, the effect on PESQ scores is minimal. However, increasing global SNR pushes more local SNRs into the range where no boosting is applied, leading to greater temporal fluctuations in the modified signal and a negative effect on PESQ scores.

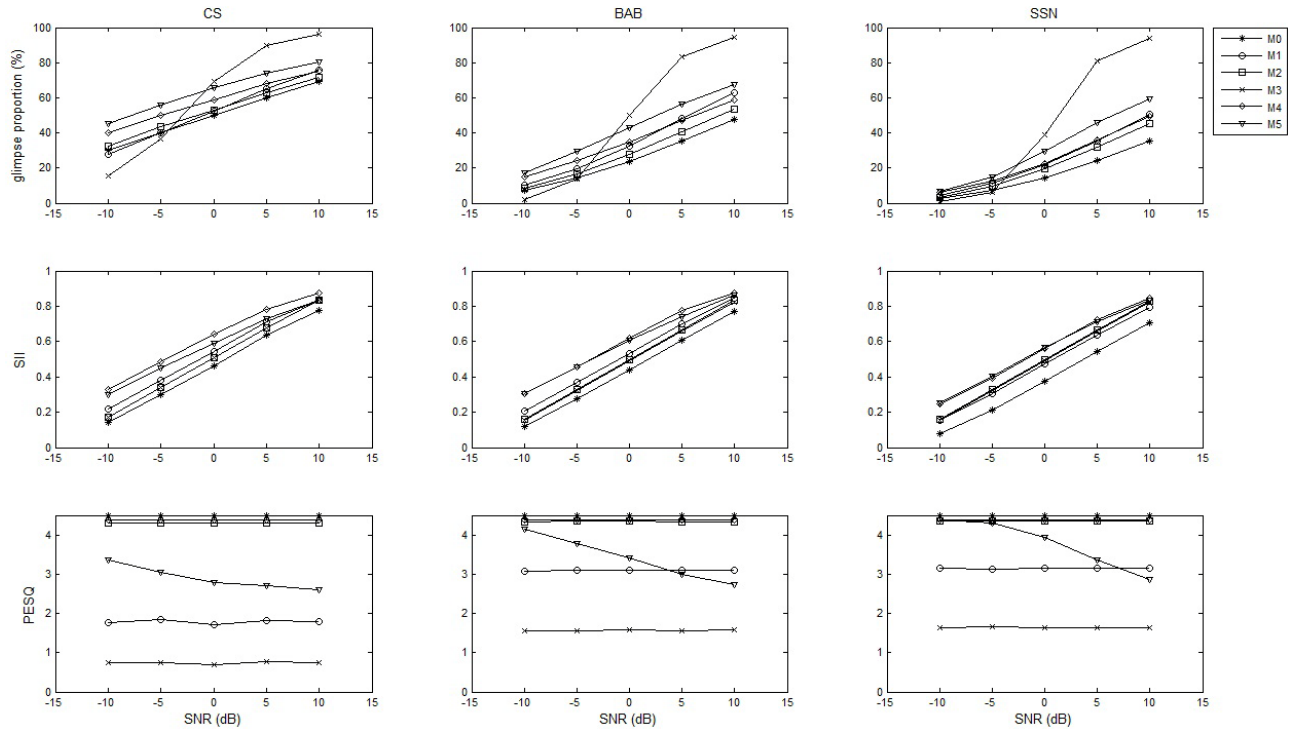

Figure 2: *Glimpse proportion, SII and PESQ as functions of SNR level for unmodified (M0) and modified speech (M1-M5) under three noise conditions (CS: competing speaker, BAB: multi-talker babbles (N=8), SSN: speech-shaped noise).*

5. Discussion and future work

Enhancing speech via energy reallocation in known noise conditions is an effective technique for intelligibility improvement. All modification strategies evaluated in the current study were predicted to improve the intelligibility of speech in noise for listeners across most SNRs and for stationary and nonstationary maskers. For instance, the M5 strategy increased objective intelligibility estimated by glimpse proportion by amounts ranging from 17% (competing speaker at 10 dB) to 147% (speech-shaped noise at -10 dB, albeit from a low baseline). However, both the degree of benefit and the effect on speech quality differ between strategies, and a tradeoff exists between the two metrics. The approaches which have the largest potential for intelligibility improvement exploit both spectral and temporal opportunities to raise the audibility of speech (M3, M5), but also cause the largest deterioration in speech quality. This is particularly the case for the attempt by M3 to maintain a constant local SNR, which inevitably results in a wide range of moment-to-moment and channel-to-channel speech energy modifications. In contrast, modification of the spectral balance alone (M2, M4) has almost no negative effect on speech quality but predicts only modest gains in intelligibility.

Notwithstanding the gains for M3 at positive SNRs, it is notable that the three equalisation strategies (M1-M3), which equate frame-, channel-, or time-frequency region-based SNRs to the global SNR, tended to underperform those approaches which employ selective amplification based on the selection of channels (M4) or epochs within selected channels (M5). The latter techniques achieve reasonable model intelligibility gains without severely compromising speech quality. The class of strategies represented by channel selection (M4) admits further development by, for example, allowing non-contiguous channels and variable boosts. Likewise, more sophisticated use of local SNR (M5) to improve audibility where it is most effective is expected to lead to further gains in intelligibility.

Since we were mainly interested in exploring baselines at this stage, all of the techniques developed here are rather crude with respect to expectations for speech quality. For instance, no attempt was made to reduce inter-channel or inter-frame gain fluctuations. A number of approaches ranging from simple post-processing via smoothing to more sophisticated application of "possible speech" constraints from articulatory models could be employed to improve PESQ scores. Use of segmental boundary information, if available (as in the case of synthetic speech), may allow masking of fluctuating energy boosts. Further, it seems likely that PESQ scores *overestimate* quality reductions caused by speech modifications since they take no account of the audibility of modified speech in noise, where poor quality may be masked. Further work will explore perceptual metrics of modified speech quality in the presence of noise. Listening tests will clarify the relationship between masked PESQ and subjective judgment of speech quality.

Practical application of the techniques introduced here will require the use of a buffer which permits a running estimate of global SNR and energy renormalisation. Live "one-way" speech applications (such as broadcast announcements) typically allow a limited delay prior to reproduction. Applications of synthetic and recorded speech containing a rich markup of, for example, segment boundaries and formant frequencies, will support a greater range of modifications.

Finally, the techniques explored in the current study were limited to those which left overall duration unchanged. In practice, speakers respond to noise by durational modifications which are segment-specific [2, 20] and at a longer-time scale by changes to rhythmic structure [1]. Exploiting this wider class of modifications can be expected to lead to further improvements in the comprehensibility of speech in noisy conditions.

Acknowledgement. This work was supported in part by the LISTA Project (http://listening-talker.org), funded from the Future and Emerging Technologies programme within the 7th Framework Programme for Research of the European Commission, FET-Open grant number 256230.

References

[1] Cooke, M. and Lu, Y., "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers", J. Acoust. Soc. Am. Conditionally accepted.

[2] Lu, Y. and Cooke, M., "Speech production modifications produced by competing talkers, babble and stationary noise", J. Acoust. Soc. Am., 124(5): 3261-3275, 2008.

[3] Bonardo, D. and Zovato, E., "Speech synthesis enhancement in noisy environments", in Proc. Interspeech, 2853-2856, 2007

[4] Yoo, S. D., Boston, J. R., El-Jaroudi, A and Li, C., "Speech signal modification to increase intelligibility in noisy environments", J. Acoust. Soc. Am., 122(2): 1138-1149, 2007.

[5] Brouckxon, H., Verhelst, W. and Schuymer, B. D., "Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments", in Proc. Interspeech, 557-560, 2008.

[6] Sauert, B. and Vary, P., "Near end listening enhancement: speech intelligibility improvement in noise environments". In Proc. IEEE ICASSP, 1: 493-496, 2006.

[7] Knobel, K. A and Sanche, T. G., "Loudness discomfort level in normal hearing individuals", Pro Fono., 18(1): 31-40, 2006.

[8] Sabin, W.E. and Schonike E.O., "HF Radio Systems & Circuits", Noble, 1998.

[9] Shannon, R. V., Zheng, F. G., Kamath, V., Wygonski, J., and Ekelid, M., "Speech recognition with primarily temporal cues", Science, 270: 303-304, 1995.

[10] Holdsworth, J., Nimmo-Smith, I., Patterson, R. and Rice. P., "Implementing a gammatone filter bank", Technical report, MRC Applied Psychology Unit, Cambridge, 1988.

[11] Strahl, S. and Mertins, A., "Analysis and design of gammatone signal models", J. Acoust. Soc. Am., 126(5): 2379-2389, 2009.

[12] American National Standard, "Methods for the Calculation of the Speech Intelligibility Index", ANSI S3.5-1997, 1997.

[13] Cooke, M., "A glimpsing model of speech perception in noise", J. Acoust. Soc. Am., 119(3): 1562-1573, 2006.

[14] Krause, J.C. and Braida, L.D., "Acoustic properties of naturally produced clear speech at normal speaking rates", J. Acoust. Soc. Am., 115(1): 362-378, 2004.

[15] Rix, A., Beerends, J., Hollier, M. and Hekstra, A., "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs", Proc, IEEE ICASSP, 2: 749-752, 2001.

[16] Loizou, P. C., "Evaluating performance of speech enhancement algorithm", Speech Enhancement: Theory and Practice, 514-525, CRC Press, 2007.

[17] Barker, J. and Cooke, M., "Modelling speaker intelligibility in noise", Speech Communication, 49: 402-417, 2007.

[18] University College London, "SCRIBE - Spoken Corpus of British English", Online: http://www.phon.ucl.ac.uk/resource/scribe/, accessed on 19 Oct 2009.

[19] Festen, J. M. and Plomp, R., "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing", J. Acoust. Soc. Am., 88(4):1725-1736, 1990.

[20] Summers, W. V., "Effects of stress and final consonant voicing on vowel articulation and formant patterns", J. Acoust. Soc. Am., 79(S1): S36-S36, 1986.