Audio Engineering Society

Convention e-Brief 637

Presented at the 149th Convention
Online, 2020 October 27-30

A Model to Predict the Impact of Dialog Enhancement or Mix Ratio on a Large Audience

Aaron Master and Hannes Müsch

Dolby Laboratories, 1275 Market Street, San Francisco, CA 94103

Correspondence should be addressed to the authors ({Aaron.Master, h.muesch}@dolby.com)

ABSTRACT

Increasing the speech-to-background mix ratio of content, either algorithmically through dialog enhancement (DE), or during production, is considered a means of reducing listening effort for an audience, some members of which have hearing impairments. But what exactly is the expected benefit? A portion of the audience can already follow the content effortlessly and dialog boosting will not improve their perception. Other parts of the audience are severely impaired, and their speech reception performance will improve until all background is removed. We introduce a model that predicts which parts of an audience benefit by how much from changing the speech-to-background mix ratio of a piece of content. The model is intended to allow decision makers to predict what impact changes in audio production guidelines or DE technologies will have on their audience.

1 Introduction

The effort needed to understand dialog in typical TV and movie content depends, among other things, on the hearing health of the listener and the speech-to-background (mix) ratio. Many listeners understand typical mixes effortlessly. For them, increasing the mix ratio will not reduce listening effort further (although some might prefer it). Other listeners have severe hearing impairments and can only understand a fraction of what is said. Their ability to understand improves until all background sounds are removed. Most of the audience will be somewhere between these extremes. But how exactly it is distributed has, to our knowledge, never been quantified. This e-Brief is a first attempt to estimate how listening effort is distributed across an audience, and how it changes as the mix ratio changes, given the audience's age and gender distribution.

The predictions draw heavily on published, peer reviewed studies. These include a study of the prevalence of hearing loss in an industrialized country [1], a model of the effect of noise and hearing loss on speech understanding (e.g.,[2,3]), and data that relate SNR to listening effort [4]. (We henceforth use SNR as a proxy for mix ratio.)

Our predictions model the effects of hearing loss and mix ratio, but these are not the only factors affecting listening effort. Listening effort is also affected by other aspects of the content, such as message complexity and speaking rate. The predictions in this brief are limited to 'everyday sentences' spoken at a 'normal' rate. Further, hearing loss is not the only dimension of the listener demographic that affects listening effort. Other dimensions, such as language proficiency (e.g., immigrant communities not listening in their first language) also affect listening effort, but they are not modelled here. Despite these limitations, we feel that the model is a valuable starting point for understanding the 'big picture' effect on an audience of mix ratio and dialog enhancement (DE, see, e.g. [5] for recent work).

In Section 2 we describe the rationale, assumptions, and steps involved in the prediction. In section 3, we illustrate the model by predicting the listening effort

distribution for two example broadcasts, each before and after a 9-dB boost in mix ratio, when the audience is assumed to be the entire US population; model limitations, sensitivity, and validation are discussed. In Section 4 we conclude and cover directions for future work.

2 Model

Our predictions are based on a simple but powerful model by Plomp ([2], see also [3]). That model predicts the Speech Reception Threshold (SRT), the lowest speech level at which listeners understand half of 'everyday' sentences. The model posits that the SRT can be predicted for any background level by knowing only: (1) the lowest level, L_0, at which a listener can understand speech in quiet and (2) the lowest speech-to-background ratio, $-\Delta L_{SN}$, at which the listener can understand speech in the presence of a high-level masker. For young, normal-hearing listeners L_0 is between 16 and 19 dB(A) and $-\Delta L_{SN}$ is about -5dB for steady maskers, becoming more negative for maskers with modulated envelopes [3]. For hearing-impaired listeners, either or both values are different from those of young, normally hearing listeners. Hearing-impaired listeners may have an *attenuation loss*, which increases L_0 and reflects that hearing loss reduces the audibility of soft speech. Hearing-impaired listeners may also have a *distortion loss*, which describes supra-threshold processing deficits that cause problems understanding even when the speech level is well above the (elevated) hearing threshold. This distortion loss increases the speech-to-background ratio, $-\Delta L_{SN}$, needed to understand, and, for many listeners, is how hearing problems first manifest.

Although it is not possible to predict any one listener's attenuation or distortion loss, we can predict attenuation and distortion loss *distributions* for a listener population given the distribution of audiograms. Consequently, our modelling begins by determining the distribution of hearing loss, as reflected in the audiogram, in the general population. For this we refer to Gablenz and Holube [1], who measured the audiograms of 1752 adults in a representative sample of two German towns and provide tables with the percentiles (0.1, 0.25, 0.5, 0.75, and 0.9) of the measured hearing loss, separately for age group and gender.[i] We condense these audiograms to a single number by taking the average of the pure-tone hearing loss at 0.5, 1, and 2 kHz (PTA3), giving us approximations of the distributions of PTA3 in that population. The PTA3 is the required input to the subsequent predictions [ii].

Next, we determine the attenuation and distortion loss distributions associated with these PTA3 distributions. Duquesnoy [3] measured the PTA3 and the SRT in quiet and at four noise levels in 110 elderly adults, used Plomp's model to find the attenuation and distortion losses that best fit the data, plotted the derived attenuation losses as a function of PTA3, and also plotted the best-fitting distortion losses as a function of attenuation loss. We read the data from these plots, fitted trend lines to predict the main effects, and modelled the residuals as normal distributions. Then, using these relations, we converted the PTA3 distributions into distributions of attenuation losses and distortion losses.

Using Plomp's model (Eq (1), the same as Eq 2 in [3]), we can now predict the SRT in dB(A) as:

$$SRT = 10 \log \left(10^{(L_o+A)/10} + 10^{(L_N-\Delta L_{SN}+D)/10}\right) \quad (1)$$

where A is the attenuation loss in dB, D is the distortion loss in dB, and L_N is the noise level in dB(A).

The SRTs predicted in this way are valid only for maskers with steady envelopes. Signals with fluctuating envelopes, e.g., music, are less-effective maskers. For young, normal-hearing listeners, the SRT is about 5 to 8 dB lower when the masker is music [6] or noise imprinted with a speech envelope [7] than when it is steady. Older listeners and those with hearing impairment are often unable to take advantage of the envelope fluctuations to the same degree. Versfeld and Dreschler [7] measured SRTs both in steady-noise maskers and maskers with fluctuating envelopes for young and old, normally hearing and hearing-impaired listeners. We use data read from their figure 6 to derive a fluctuating-masker 'bonus' to be applied (only in cases of such maskers) to the SRTs computed in the previous step, taking into consideration PTA3 and age. As was the

case for the previous transformations, the relation between input and output is probabilistic, with the parameters of the probability distribution estimated from the data.

We considered two alternative methods to interpret the predicted SRTs. One estimates the ability to understand, and the other, the effort needed to understand. To estimate the ability to understand, we recall that the SRT is the speech level at which, at a given background level, a listener understands half of what is said. Generally, the speech level of content will be significantly higher, so we must estimate how speech recognition changes as the speech level moves away from SRT. That relation is described by the Performance-Intensity function (PI), a sigmoid-shaped function that relates the speech level, here expressed relative to the SRT, to the proportion of speech understood. For our predictions we use the PI functions reported in [8], as these functions were measured under conditions matching those assumed in the predictions made above.[iii] When speech level is expressed relative to the SRT, the PI functions for normal-hearing and hearing-impaired listeners intersect at the 50%-correct point, by definition.

Expressing results as listening effort is appealing because it allows differentiating between conditions with very good intelligibility. To predict listening effort, we use figure 7 of [4], which shows normal-hearing listeners' rating of listening effort to be a linear function of SNR. To predict the effort rating for a hearing-impaired listener, we first use the PI function for the hearing-impaired listener to look up the predicted speech recognition performance, then use the PI function for young normal-hearing listeners to find the speech level needed to achieve that same performance, and calculate the difference in speech levels. We then offset the actual SNR in the content by that difference and read the associated effort rating. This approach implements the assumption that, while normally hearing and hearing-impaired listeners require different SNRs to achieve a given recognition rate, they must put in the same listening effort to achieve that rate[iv]. The steps outlined up to here are universal and would be applied in any prediction. The steps that follow, i.e., the selection of content to be analyzed and the description of the audience, will be adapted by the user to their unique circumstance.

3 Results

We now use the model to predict specific audience reception of two sets of sentence-length phrases, sampled from broadcasts, each before and after a 9-dB boost in mix ratio (as could occur from DE). The first set was drawn from three sports broadcasts with commentary over crowd noise and included phrases with mix ratios from 4.5 to 10 dB; the second set was taken from two documentaries, where narration was underlaid with music, and mix ratios were from 3 to 7.5 dB. For each set, we assumed a reproduction level of 58 dB(A), which [9] found to be the mean preferred listening level of home television viewing. The choice of listening level is not critical for the predictions of this study, however, as any intelligibility problems stem almost exclusively from the low mix ratio rather than insufficient reproduction level.

We chose this content because it was real broadcast material that differed in background type, allowing us to illustrate the model with and without the 'fluctuating noise bonus' described above. The crowd noise of the sports broadcasts was steady over the duration of a phrase or varied only slowly, so we did not apply the bonus there. The background music for the documentary met the "fluctuating" definition so the bonus was applied. (The model will benefit from a more formal way to decide on the degree to which masker envelope fluctuations are present in the program material.) We then applied the model repeatedly, once for each phrase, to compute the SRT distribution across gender and age group, and, observing the corresponding speech level, computed the distribution of listening effort.

The results of these predictions are shown in figures 1 and 2. The color coding of each horizontal bar represents the distribution of effort ratings for a gender and age group; females are on the left and males on the right, with age groups listed at left. The length of the bars is scaled to reflect each group's size relative to the total US population, making the graph resemble a population pyramid. Inspection of

the plots shows that most of the younger population understands with little to no effort, and that the proportion of listeners needing to exert more than just a little effort increases with age. A 9-dB boost in speech level (via changed mixing or DE) leads to a substantial reduction in listening effort, except for listeners who already understood with no effort. For the documentary, we observe relatively low effort levels before dialog boost, due to the fluctuating noise bonus. For the boosted version, the bonus remains reduced for the older hearing-impaired population as noted above, leading to more limited improvement for this population.

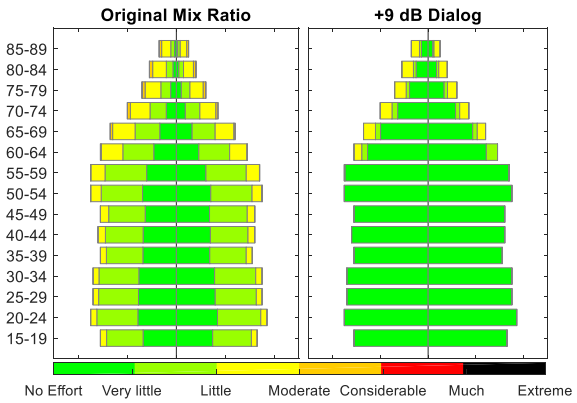

Fig. 1. Predicted effort ratings for sports broadcast.

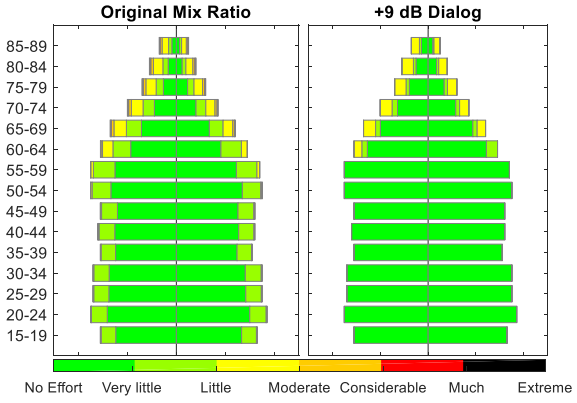

Fig. 2. Predicted effort ratings for documentary.

A model with as broad a scope as this is difficult to validate. We can, however, assess the impact of estimation errors in the probabilistic transforms used to derive SRT estimates from PTA3 distributions. When estimating transforms from the data, we obtained 95% confidence intervals for the estimated parameters. Figs. 1 and 2 show predictions when the best estimates are used. Fig. 3 shows predictions for the original sports content when we *always* use the 'pessimistic' (predicting low effort) end of the confidence interval (left) or when we *always* use the 'optimistic' end (right). The pattern of results is unaffected (roughly a shift of one point on the effort scale) as are conclusions about the effect of speech level boost (not shown).

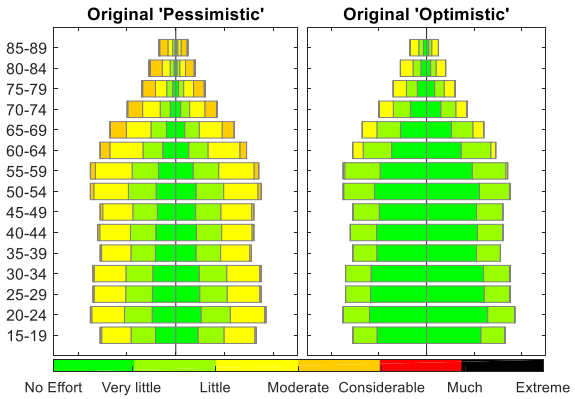

Fig. 3. 'Pessimistic' (left) and 'optimistic' (right) predictions for the sports content as broadcast. (The unbiased estimate is in the left panel of Fig. 1.)

4 Conclusions and future work

Though this work does not capture all aspects of speech understanding, we believe that the model already allows a glimpse of the return on investment that one can expect from implementing dialog enhancement technologies in a delivery chain.

Nonetheless, we intend to expand the model to include the effect of non-native speech perception; the effect of noise on this group has been measured extensively, and we expect implementers could estimate its proportion of an audience. Also not modelled is the effect of spatial presentation, which varies across home listening environments. It is generally accepted that spatially separating talkers from maskers can lower the SRT, thereby lowering effort vs that for a downmix. Impact and population relevance of this should be investigated in the future.

References

[1] P. von Gablenz and I. Holube, "Hearing threshold distribution and effect of screening in a population-based German sample" *Int'l Journal of Audiology* 55, 110–125. (2016).

[2] R. Plomp, "Auditory handicap of hearing impairment and the limited benefit of hearing aids" *The Journal of the Acoustical Society of America* 63, 533; (1978).

[3] J. Duquesnoy, "The intelligibility of sentences in quiet and in noise in aged listeners," *The Journal of the Acoustical Society of America* 74, 1136 (1983).

[4] J. Rennies, H. Schepker, I. Holube, and B. Kollmeier, "Listening effort and speech intelligibility in listening situations affected by noise and reverberation." *The Journal of the Acoustical Society of America* 136, 2642 (2014)

[5] A. Master, L. Lu, H.-M. Lehtonen, H. Mundt, H. Purnhagen, and D. Darcy, "Dialog Enhancement via Spatio-Level Filtering and Classification," in *AES 149th Convention*, New York, 2020.

[6] D. Başkent, Engelshoven, and Galvin III "Susceptibility to interference by music and speech maskers in middle-aged adults." *The Journal of the Acoustical Society of America*, 135, EL147 (2014)

[7] N. Versfeld and W. Dreschler "The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners." *The Journal of the Acoustical Society of America*, 111, 1; (2002)

[8] J. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing" *The Journal of the Acoustical Society of America*, 88, 1725 (1990)

[9] E. Benjamin, "Preferred Listening Levels and Acceptance Windows for Dialog Reproduction in the Domestic Environment" *Audio Engineering Society*. Convention Paper 6233 -- 117th Convention. (2004).

[10] ISO 7029: 2017, "Statistical distribution of hearing thresholds related to age and gender"

[11] G. Schmoorenburg, "Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram" *The Journal of the Acoustical Society of America*, 91, 421 (1992)

[12] I. Holube, K. Haeder, C. Imbery, and R. Weber, "Subjective Listening Effort and Electrodermal Activity in Listening Situations with Reverberation and Noise," *Trends in Hearing* 20: 1-15 (2016)

[i] An alternative, but we feel less representative, source is ISO 7029 [10], which tabulates the distribution of audiograms for otologically normal men and women as a function of age. "Otologically normal" means that only the natural reduction in hearing ability with age is described and individuals with noise-induced hearing loss are excluded. ISO 7029 is based in part on [1]. To derive the subset of otologically normal individuals, 819 of the 1752 participants (47%) in [1] had to be excluded. This emphasizes the need to use a representative sample, such as [1], instead of ISO 7029.

[ii] There is evidence that including the hearing loss at higher frequencies provides additional predictive power (see e.g.,[11]). However, those data have not been incorporated into our model at this early iteration.

[iii] The shape of the PI function depends on several factors, including the speech material (reflecting that highly redundant messages, e.g., proverbs, can be understood more easily than less-predictable messages), the language proficiency of the listener, and the spectral match between speech and masker. The PI functions we use are for 'everyday' sentences in the listener's native language.

[iv] This assumption warrants further investigation. While it is frequently argued that older and hearing-impaired listeners need higher effort to achieve the same performance as young normally hearing listeners, other data (e.g.,[12]) support our assumption.