Hearing Enhancement

By Francis Rumsey
Staff Technical Writer

There is a growing awareness of the problems of intelligibility that arise for people with hearing impairments when they listen to reproduced sound. Difficulty understanding the content of spoken material in particular can affect both live performance situations involving PA systems and domestic listening situations. It naturally affects the elderly more than the rest of the population, because the incidence of hearing impairment is greater in this sector, but it is by no means irrelevant to many other groups that may suffer hearing difficulties of one kind or another. In this article a number of recent AES convention papers relating to hearing enhancement are summarized.

ASSISTIVE AUDIO SYSTEMS FOR THE HARD OF HEARING

Mapp, in "Assessing the Acoustic Performance and Potential Intelligibility of Assistive Audio Systems for Hard of Hearing and Other Users" (AES 125th Convention paper 7626), considers reasons why the increasing number of assistive hearing systems (AHS) being installed may fail to perform adequately in practice. He explains that roughly 14% of the population suffers from noticeable hearing loss and that some 4 to 6 million people in the UK would benefit from the use of a hearing aid. There are actually about 2 million regular wearers of such aids in the UK. Data suggest that the situation is similar in other countries. However, it appears that surveys indicate assistive systems to be working poorly or not at all in many cases. One clear problem is that no standards and little guidance are available to help in the testing of such systems for intelligibility, which is needed in order to demonstrate that intelligibility is adequate. (There is an IEC standard, 60118-4, designed for testing the electronic and electromagnetic performance of inductive loops, but this does not deal with acoustical issues such as mentioned here.) Mapp goes on to review different components of such systems in an attempt to highlight areas for improvement and to suggest a way of ranking systems according to a simple scheme.

He shows that while audio frequency induction loop systems (AFILS) are probably the most commonly encountered, infrared and FM wireless systems are also used. Most UK hearing aids have induction loop settings, although the frequency range that they handle adequately tends to be from around 500 Hz to 5 kHz, while the other two systems do not suffer from such limited bandwidth. In the USA and Europe there is a wide disparity in the use of AFILS, both in hearing aids and in installations. As shown in Fig. 1, all systems have common components and features of the signal path, including a means of picking up the sound (affected by reverberation and noise), a transmission system (affected by noise), and a receiver system (affected by frequency response limitations).

One of the main factors to address in such systems is the source pickup by microphones. If this is derived from close, body-worn microphones mixed for the PA system then the direct-to-reverberant ratio is likely to be high enough to achieve good intelligibility. However, many venues employ general coverage microphones mounted on a ceiling or a lighting gantry at some distance from the sources, which can give rise to unacceptable levels of reverberation. The graph in Fig. 2 shows how the speech transmission index (STI) is affected by distance from the source for different types of microphones in a hall with two seconds reverberation time. He found that even when two nominally cardioid microphones were compared, they gave rise to

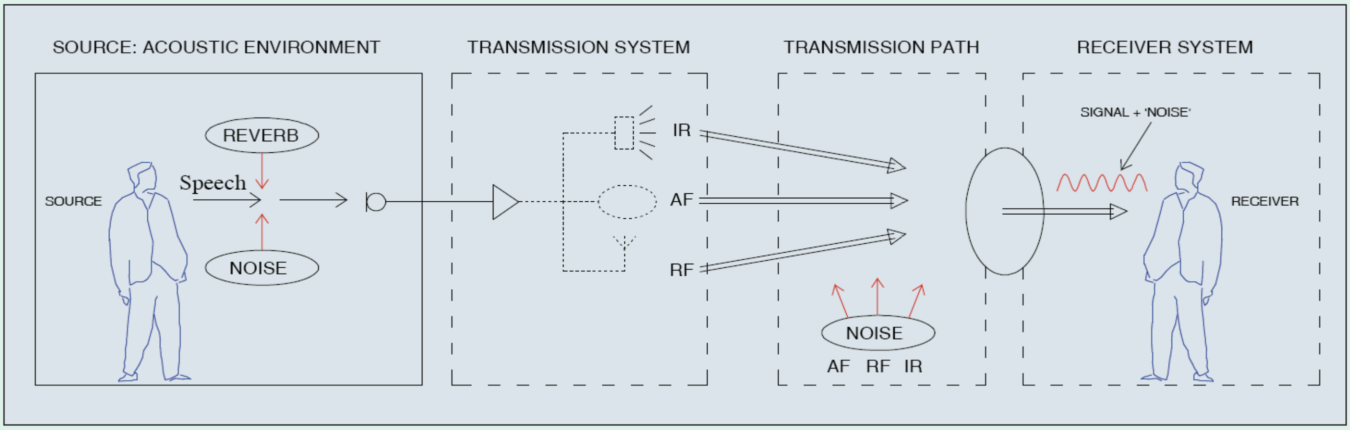

Fig. 1. Assistive hearing system transmission chain showing sources of interference (Figs. 1–3 courtesy of Mapp)

noticeably different STI results owing to the differences in their off-axis frequency responses. When the feed for the AHS is derived from the show relay pickup (used to feed backstage areas and dressing rooms in a theater) this is rarely tested for its suitability to help hearing-impaired listeners. For these reasons and others Mapp suggests ways of testing such systems using portable direct-reading STIPa (STI for PA systems) equipment.

Mapp also points out that the artificial talker used to measure the STI in systems of this type needs to be representative of a real human talker in terms of its directivity. Although IEC 268-16 suggests that a 100-mm driver in a small enclosure can be used in this context, this does not do the job adequately and can lead to noticeable errors in the measurement of intelligibility. This is partly because of the reflections that are stimulated in different parts of a hall as a result of varying talker directivity. For this reason the author has chosen to develop his own talker with more suitable characteristics, the directivity plot of which at 2 kHz is shown in Fig. 3.

Background electromagnetic noise can add to the problems of intelligibility in AFILS systems, such as noise from fluorescent lights, from power systems, and from signalling systems in railway installations. However, it seems that much of the low-frequency component of such noise is filtered out as a result of the inductive pickup response of most analogue hearing aids. Mapp suggests an "H" frequency response curve to weight the noise measurement accordingly, which emulates this characteristic reasonably closely. He concedes, however, that the standard A-weighting curve may be more suited to the current generation of digital hearing aids that have a T-coil fitted, and which have a better frequency response. It is apparent that the limited frequency range of many basic hearing aids means that the best STI achievable is only 0.7, which means there is little margin for additional reductions due to other features of a sound system if a target criterion of at least 0.6 is aimed for. He proposes a set of bands for STI in AHS applications, running from A+ to I (the lowest for useful PA), which can guide those attempting to evaluate the suitability of a system for a particular installation. The highest band relates to communication systems, with those in the middle relating to settings such as shopping malls.

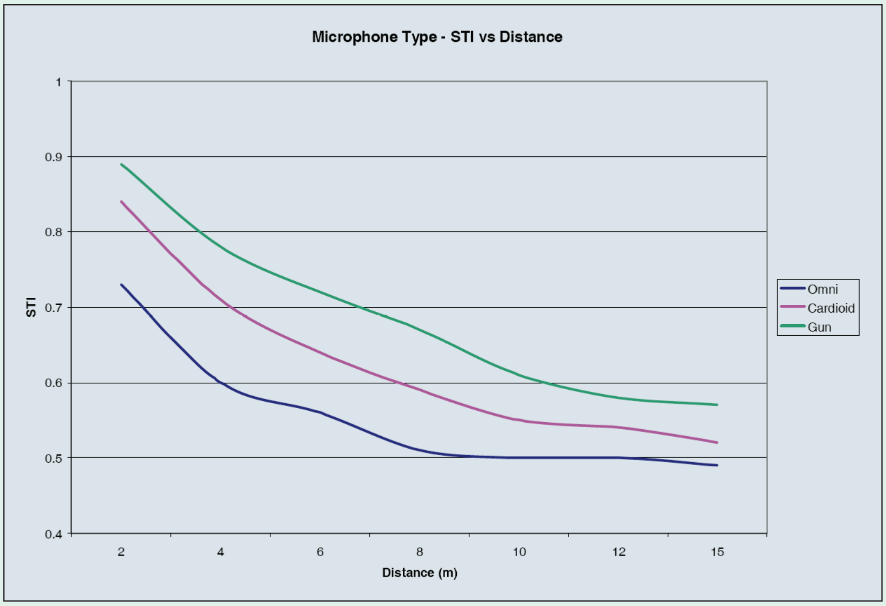

Fig. 2. STI versus distance for a 2-second RT recital hall comparing omni, cardioid, and gun microphones

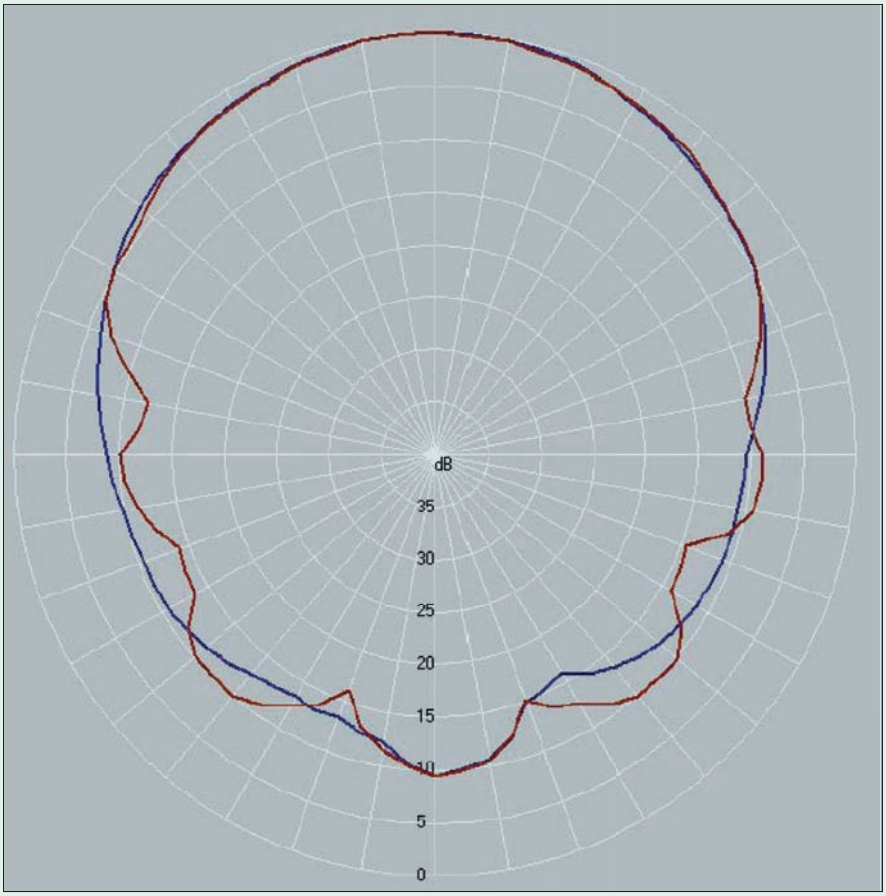

Fig. 3. Directivity plot of Mapp's experimental talker at 2 kHz

"WHAT WAS HE SAYING?"

Even those with normal hearing sometimes find it hard to understand the dialogue in movies, so it comes as no surprise to discover that people with hearing impairments struggle mightily in this regard. The presence of high-

STIPa is a version of the speech transmission index (STI) intended for PA systems. It is a stripped-down version of STI that provides good evaluation of the effects of time-domain and band-limitation distortions, but limited evaluation of the effects of nonlinear distortions. STI is measured on a scale from 0 to 1, where 0 represents completely unintelligible and 1 represents perfectly intelligible. Mapp suggests that a value of 0.6 or higher is typical of a good concert hall or reasonable theater sound system. STI works by measuring the signal modulation characteristics in a number of frequency bands across the audio range. If the modulation depth of the test signal is reduced because of noise, interference, or reverberation, then it is assumed that intelligibility is worsened.

level background music and effects serves to make speech difficult to understand in many cases. And television sound can suffer from similar problems. The intelligibility of speech is the single most important factor, putting niceties such as surround sound and the ability to hear subtle effects in the shade.

According to data presented by Müsch in "Aging and Sound Perception: Desirable Characteristics of Entertainment Audio for the Elderly" (AES 125th paper 7627), 65 is the median age of the audience for the U.S.'s *Fox News*. For the *Late Show with David Letterman* it is 53. There is a clear trend toward older listeners and viewers. As shown in Fig. 4, the reduced hearing thresholds of older listeners tend to result in the loss of quite a large part of the speech spectrum. These listeners also tend to prefer higher listening levels than younger listeners, although because of loudness recruitment the level at which listening becomes uncomfortable is similar, so there is reduced window of comfortable dynamic range. For these reasons and others, Müsch goes on to identify the important factors governing the comprehension of speech by elderly listeners and isolates those that can be improved by signal processing.

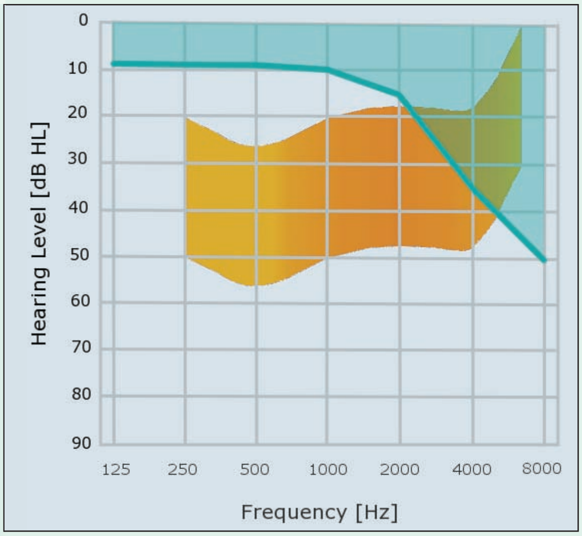

Fig. 4. Median hearing threshold of 65-year-old men in relation to the level and frequency range of speech sounds (shown dark shaded, Figs. 4–6 courtesy Müsch)

He discovers that providing a "clean" sensory input is crucial, because otherwise too much mental processing is allocated to repairing the information and too little power or attention is available for the higher level cognitive processing that places words in the context of surrounding sentences (which can aid comprehension and the filling in of missed words). It seems that elderly listeners' ability to use working memory to help with the cognitive process of comprehension is not particularly impaired, rather their sensory ability is mainly responsible for any difficulties. The speech intelligibility index (SII) is a very useful measure for understanding the parts of the spectrum that contribute most to comprehension. As shown in Fig. 5, the band importance peaks between 2 and 4 kHz. This is combined with the level weighting function that multiplies the spectral band weight according to how much energy is present in the speech signal, compared with the hearing threshold.

Temporal factors also seem to play a part in speech comprehension, and it is explained that older people have greater difficulty detecting the short gaps in speech that can distinguish one word from another, particularly in noise. This is thought to be related to a deterioration in the phase-locking ability of hair cells in the inner ear. There is also known to be a gradual widening in the bandwidth of the auditory filters, which tends to smear the spectral

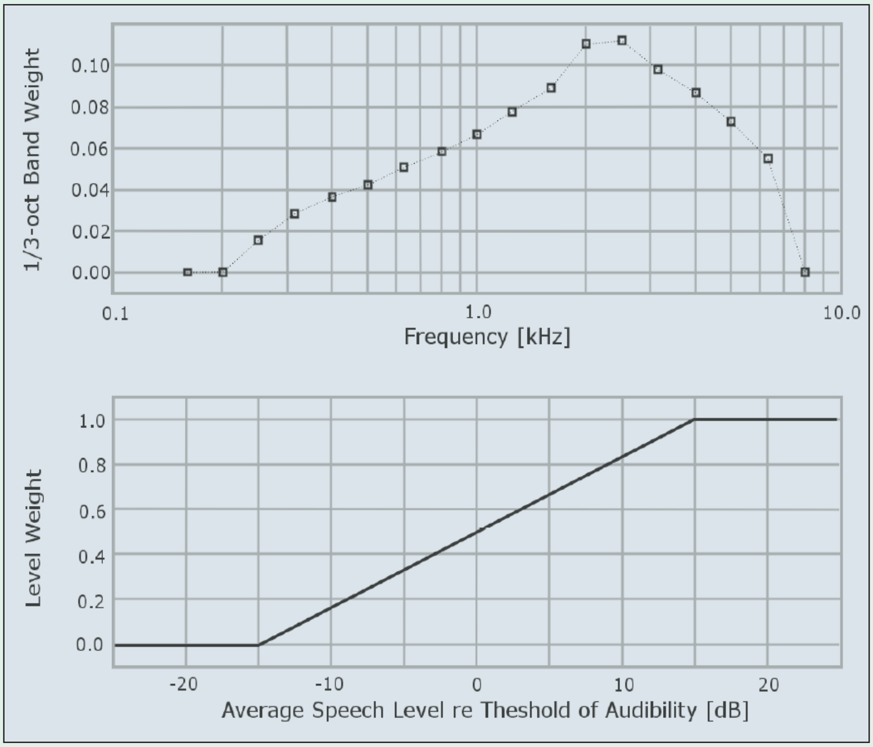

Fig. 5. Top, the frequency importance for meaningless syllables as defined by the speech intelligibility index; bottom, band importance as a function of level

shape encoding of speech cues (such as the formant pattern of a voice).

"Clear speech," says Müsch, proceeds at about half the rate of conversational speech, it also requires longer plosives (p and b sounds, for example), higher level consonants, and a larger difference between vowel formant patterns. One might therefore expect that slowing speech using time-stretch processing might aid comprehension. But most research conducted to date suggests this is not successful and neither does it seem to be related to any processing artifacts. The reasons for this are not entirely clear. Another option, spectral contrast enhancement, seemed to have shown mixed results in the literature, ranging from modest improvement to reduced intelligibility. The author concludes that speech audibility is the single most important factor governing good speech reception, and proceeds to describe an algorithm to improve this in television audio.

One of the design criteria was that the processed signal should be immediately preferred by listeners, which is different from the situation with hearing aids, where many users dislike the result at first. Furthermore, the processing should not have to be tuned to individual requirements. The algorithm was designed to treat the two main reasons for lack of clarity: parts of the speech are masked by other content and parts of the speech are below the hearing threshold. As shown in Fig. 6, the author takes advantage of the fact that most 5-channel surround mixes carry dialogue primarily in the center channel. For this reason the other nonspeech channels are selectively "ducked" (reduced in level) when their spectral content would appear to reduce the comprehension of speech in the center channel. The speech intelligibility index, mentioned earlier, is used as a guide. However, the center channel does not always carry speech, so the likelihood of speech presence has to be detected by another algorithm. In order to ensure speech audibility when it is not masked by other signals, a multiband wide-range compressor is used to process the speech signal, resulting in a compression ratio of no more than 2 to 1 and a gain of 10 dB.

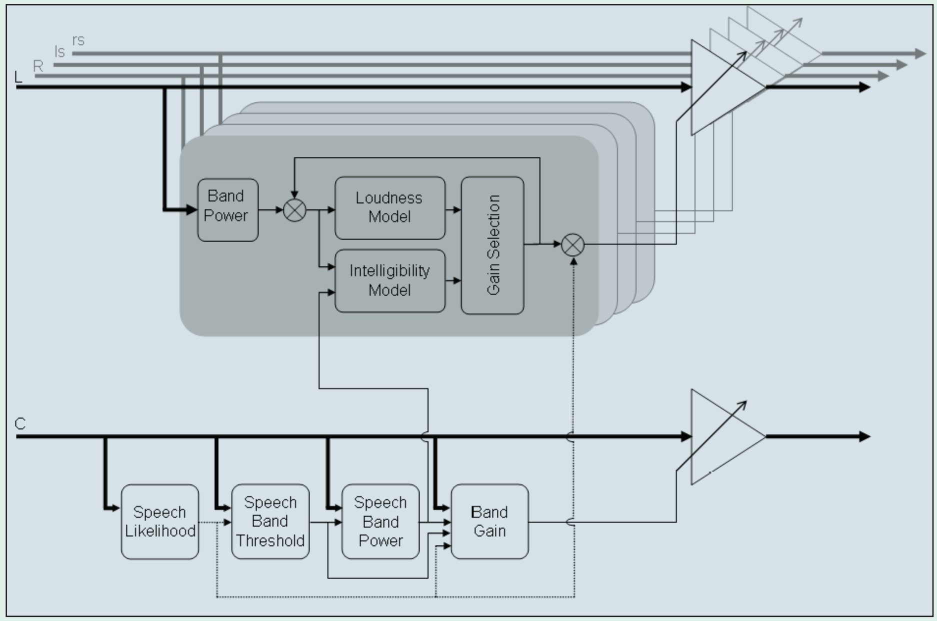

Fig. 6. Flow graph of an algorithm to enhance speech audibility in a 5-channel entertainment program. Heavy lines represent audio signals, such as blocks of 256-point MDCT transform coefficients. Solid thin lines represent *n*-dimensional control signals (powers and gains) that correspond to *n* frequency regions of the signal. The dotted line represents the likelihood of speech presence, which is a scalar. Note the two interacting side-branch structures. The four signals at the top (L, R, ls, rs) represent nonspeech channels. The variable gain blocks associated with these signals represent four banks of *n*-band attenuators. The single signal at the bottom (C) represents the speech channel. The variable gain block represents an *n*-band wide-dynamic range compressor.

Although Müsch says it is not expected that his algorithm will improve individual word recognition, he hopes that it will reduce the listening effort involved and thereby improve comprehension and enjoyment of listening. He suggests that there do not appear to be accepted metrics to quantify ease of listening at present.

SPEECH DETECTION AND ENHANCEMENT

Uhle et al. tackle a similar problem in a slightly different way, which is described in "Speech Enhancement of Movie Sound" (AES 125th paper 7628). Like Müsch, they want speech to be enhanced in the presence of interfering music and effects, so that people with hearing impairments can follow the plot of a movie. They also say that those without hearing impairments may find this useful as well, because late-night listening is often a compromise between being able to follow the plot and annoying the neighbors or the rest of the family. For these reasons they developed an algorithm that detects "relevant" speech and then enhances it while attenuating other sounds. (By relevant, they mean speech that is necessary to follow the plot of a movie, so a crowd of people talking together is regarded as noise for example.) Using this algorithm nonspeech components are attenuated by 12 dB or more and the processed speech signal can be rendered at a higher output level because the loud effects sounds and music are quieter.

The speech-detection algorithm uses a pattern-recognition system shown in Fig. 7, which employs a wide range of features to identify relevant speech signals. The authors chose to concentrate on monophonic algorithms partly because there is a lot of legacy movie content in mono, and also because they, like Müsch, work only on the center channel of a surround mix. They also suggest that the audio channels of a stereo movie are often highly correlated. An average recognition rate for speech of nearly 90% was obtained, which is lower than the results from other work known to the authors. However, they say that there are particular characteristics of movie sound that make speech detection

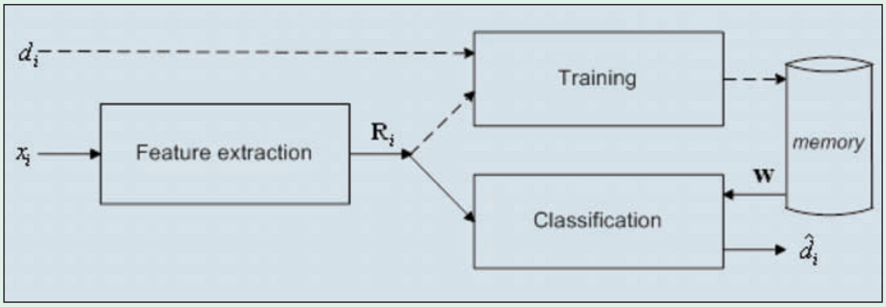

Fig. 7. Block diagram of speech pattern-recognition system for use with movie sound. Objects x_i are classified using a set of features $\mathbf{R}_i$. The parameter set **w** of the classifier is obtained during a training process (shown by dotted lines) from examples with known labels d_i and are stored in memory. The parameters are then used to classify novel patterns (shown by solid lines; Figs. 7 and 8 courtesy Uhle et al.).

difficult. These include the variety of speech sounds and the high background noise levels typically encountered. One solution they suggest is to use speech enhancement for slightly longer than the detected speech regions, in cases where the segment boundaries between speech and nonspeech are not accurately detected.

The speech enhancement algorithm used spectral weighting and noise estimation based on a supervised learning method, splitting the signal into a number of subbands. This can deal with a range of nonstationary background sound using feature extraction and a neural network regression method. The processing method proposed by the authors (PM) was compared with two previous noise estimation algorithms (C1 and C2). When evaluating these algorithms, listeners were asked to evaluate sound quality and speech quality, as well as noise reduction. The listeners fell into two groups: hearing impaired children and normal hearing listeners. The former group rated the sound quality of unprocessed audio, C1, and PM as equally good, and higher than that of C2, while the latter group preferred the unprocessed audio. Speech quality ratings were similar in pattern, with unprocessed, C1, and PM being insignificantly different for the hearing impaired listeners. Noise reduction was, however, noticeably greater with PM than with the other methods.

AUDIO BALANCE FOR THE ELDERLY

The best sound balance for elderly listeners has been studied recently by Komori et al. and discussed in "An Investigation of Audio Balance for Elderly Listeners Using Loudness as the Main Parameter" (AES 125th paper 7629). They had already observed that elderly listeners benefited from a reduction in background sound level (nonspeech sound) by 6 dB compared with that used in normal broadcast sound balances. They note, however, that this balance is determined subjectively by sound engineers so it is hard to find an objective reference point. Consequently they undertook subjective tests with professional sound engineers, comparing the balance between narration and background sounds and analyzing the results in terms of loudness.

Because the traditional VU meters used in Japanese sound broadcasting are a relatively poor indicator of loudness, the authors employed loudness measurement using the standard objective calculation methods specified in ISO 532B and ISO 226. Although ITU-R BS.1770-1 discusses algorithms for measuring long-term loudness of program material, the authors decided to use a short-term loudness calculation method because the speech rate concerned seemed to warrant a real-time measure. For this reason they measured loudness in 42-ms frames with 50% overlap, although only taking into account those frames with high loudness levels (this is because other research suggested that perceived program loudness depends mainly on those periods in which the loudness exceeds a certain threshold). Because most consumer television sets have poor low-frequency response, they ignored bands below 250 Hz. They called the resulting measure "weighted short-term loudness" and the mean value of terms with high levels of this "weighted loudness."

During the experiment, experienced sound engineers were asked to adjust the background sound level of documentary program material relative to the narration in order to achieve an "appropriate sound balance for broadcast." The average background sound level was then measured in relation to the narration, and the resulting short-term weighted loudness is shown in Fig. 9. The average difference between the two was approximately 9 ± 3 phons. In a separate experiment, the ability of elderly listeners to identify monosyllabic words in the presence of background noise was evaluated and it was found that there was a significant difference in their ability compared with that of young listeners at some relative levels, although both groups suffered similar impairments because of the noise.

The investigators made the assumption that elderly listeners suffer from a higher degree of masking of speech by background noise, and from recruitment, as mentioned above. To study this

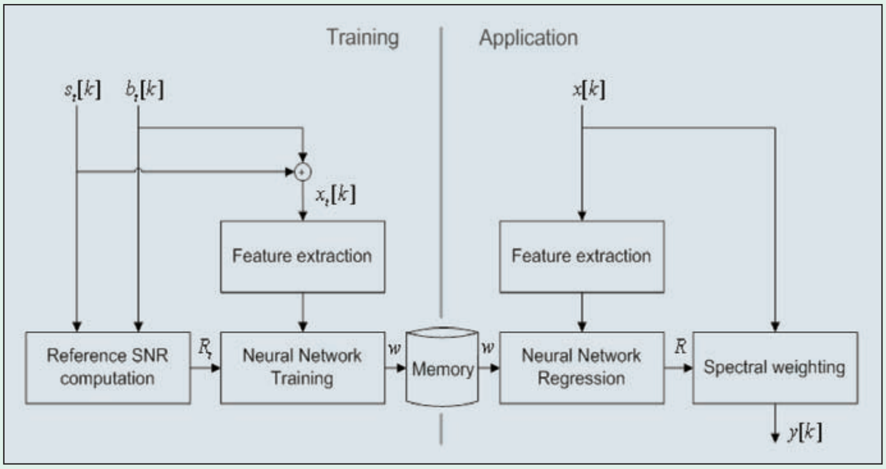

Fig. 8. Spectral weighting processing used by Uhle et al. for movie speech enhancement

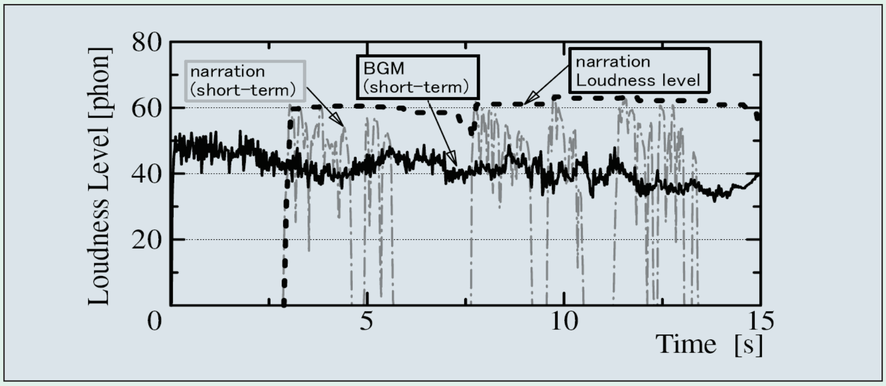

Fig. 9. Short-term weighted loudness of narration and background sounds in documentary programs balanced by sound engineers (Figs. 9 and 10 courtesy Komori et al.)

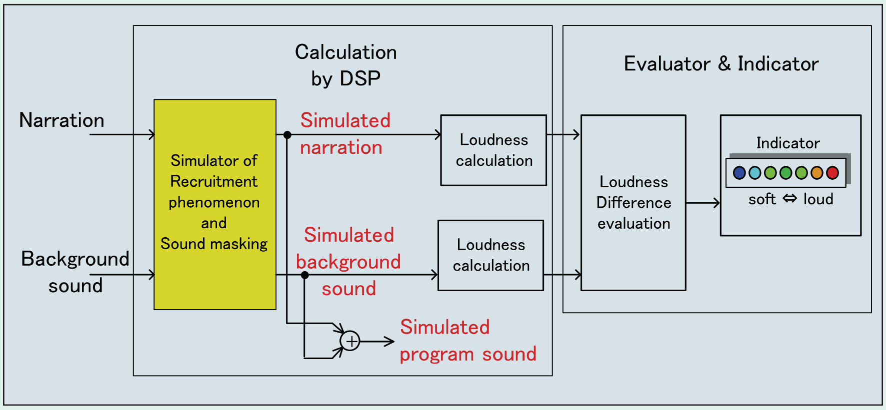

Fig. 10. Block diagram of prototype system for evaluating relative loudness of narration and background sounds incorporating simulation of hearing impairment

they developed a signal processor that simulated the recruitment phenomenon and increased masking effect noticed in elderly listeners. They used this to develop a meter that could indicate the likely perceived loudness difference for such listeners on a simple scale, as shown in Fig. 10. This clearly requires separate inputs from narration and background audio streams. In the future they plan to use such a system to collect experimental data that may help to develop support technologies and to guide the production of high-quality broadcast sound.

SUMMARY

The authors of the papers described here have all been concerned to improve the experience of listening to entertainment and public audio for people with hearing impairments. However, it seems clear from what they found out that the situation can be far from ideal for normal hearing listeners too, and that methods employed to assist the hard-of-hearing may also prove beneficial for others. Clearly perceived speech is the overriding factor of importance for everyone, which is not surprising considering its importance for basic communication. The effects so beloved by movie makers, while increasing enjoyment in some respects, are often detrimental to understanding of plot and narrative. The solution is either some form of effective monitoring and control at the mixing stage or downstream processing at the receiver end to detect and correct any problems that are noticed.

Editor's note: The papers reviewed in this article, and all AES papers, can be purchased online at <www.aes.org/publications/preprints/search.cfm> and <www.aes.org/journal/search.cfm>. AES members also have free access to past technical review articles such as this one and other tutorials from AES conventions and conferences at <www.aes.org/tutorials/>.