Research White Paper

WHP 190

April 2011

Audio Processing and Speech Intelligibility:

a literature review

Mike Armstrong

BRITISH BROADCASTING CORPORATION

BBC Research & Development
White Paper WHP 190

Audio Processing and Speech Intelligibility: a literature review

Mike Armstrong

Abstract

This paper is a literature review on the subject of audio processing and intelligibility. It looks at the problem of extracting speech from noise and reviews the success of such techniques in improving intelligibility in a number of fields of research. The literature available indicates that there is little if any chance of audio processing improving intelligibility of speech in noise, and a real danger of degrading it. Whilst audio processing can be used to create cosmetic improvements in a speech signal it cannot be used to improve the ability of an audience to follow the words.

Audio processing cannot be used to create a viable “clean audio” version for a television audience and any use of noise reduction behind speech in production will have similar problems.

Additional key words: signal separation, source separation, unmixing, independent component analysis.

White Papers are distributed freely on request.

Authorisation of the Head of Research is required for publication.

Audio Processing and Speech Intelligibility: a literature review

Mike Armstrong

1 Introduction

This White Paper is a review of published research in the area of audio processing to remove background sounds. It is known that background sounds reduce the ability to comprehend speech and that this effect is greater for those who have some form of hearing loss. For broadcasters, a tension can arise between the desire to create rich multilayered soundtracks and the desire of some sections of the audience to hear the speech without accompanying music or sound effects. It has been suggested that broadcasters could provide an alternative "clean audio" mix, but the production of a second soundtrack is time consuming and expensive, and does not address the majority of problems audiences have with broadcast speech [1]. It has been suggested that the alternative might be to create a cleaned-up soundtrack by processing the normal audio to reduce the level of background sounds and improve the intelligibility of the programme [2].

This paper brings together published work in the area of sound processing to reduce background noise behind speech and its impact on intelligibility. In all cases there is no clear evidence that speech processing can be used to reliably improve intelligibility. Indeed, in many cases the processing could reduce the intelligibility of speech.

2 Background

In 2010 BBC Vision carried out two separate surveys with its *Pulse* on-line panel of 20,000 viewers to identify the issues which caused problems with television sound. The research showed that nearly 60% of viewers had some trouble hearing what was said in TV programmes [3].

The research identified four key factors that can make it hard for viewers to hear what is being said:

- Clarity of speech: poor and very fast delivery, mumbling and muffled dialogue, turning away from camera, people talking over each other, trailing off at the end of sentences.
- Unfamiliar or strong accents: Audiences find accents other than their own harder to understand.
- Background noise - locations with heavy traffic, babbling streams, farmyard animals, in fact any intrusive background noise can make it difficult to hear what's being said.
- Background music - particularly heavily percussive music or music with spikes that cut across dialogue.

Any of these issues can create problems for viewers, but the research showed that when these factors combine, then many people struggle to understand [1].

As part of the study the Voice of the Listener & Viewer gathered diaries from another 506 people aged 65 and over who did not use the internet. The VLV give figures for the response across 22 "problematical" programmes. They report that 11% of the respondents cited problems with background music and 13% with background noise whilst 19% had problems with accents & dialects, 14% with mumbling and poor diction and 11% with talking too fast [4]. These figures suggest that the removal of music and sound effects to create a "clean audio" soundtrack could, at best, improve the audibility of the speech in only a quarter of cases. New BBC Editorial Policy Guidelines [1] have been produced alongside a series of online training videos setting out best practice in creating clear sound [5].

3 Audio Processing and the Signal Separation Problem

Background noise reduction is a form of signal separation problem; two audio signals can be separated completely if, and only if, they are in some way orthogonal. The simplest example is where you have two different linear mixtures of two individual sound sources. In this case the sources can be separated by further linear mixing of the two signals to arrive back at the original sources. The most difficult part is working out how to recombine the mixtures to arrive back at the original signals. The problem increases considerably for three or more signals containing three or more sources. This linear separation can only work if there have been no further changes to the signals, any reverberation or distortion will make this process impossible, resulting in residual noise signals which cannot be removed from the individual sources [6].

In the case of a speech signal which is mixed in with background noise or music we do not have sufficient linear mixtures to separate out the speech using linear unmixing. However, there are two cases where we can obviously achieve orthogonality by other means, this is where the two signals are either separated in time or separated in frequency. If the noise does not occur during the speech then it is relatively straightforward to remove the noise by blanking out those portions of the signal where the noise occurs. Likewise if the speech and the noise do not overlap in frequency then the two signals can be separated with a high pass and low pass filter, e.g. the removal of 50Hz hum underneath speech.

If the speech and noise overlap then some form of modelling of the two signals is needed to ensure good separation. In simple cases, for example if speech has been mixed with 1kHz tone which overlaps the speech in time and frequency, it is possible to detect the phase, level and frequency of the tone enabling the tone to be synthesised accurately enough to subtract it from the speech. However, if the tone was varying in pitch or level then it would only be possible to remove the tone with a notch filter. Whilst this could completely remove the tone it would also remove a band of frequencies from the speech. The result may be more plesant to listen to than with the tone, but the residual damage left by the filtering cannot be repaired.

The general case of separating speech from background sound is considerably harder. The competing signal may be relatively static in character, such as white noise, or dynamic noise such as the sound of passing traffic or the interior of a pub or restaurant. With static noise there is a reasonable chance of characterising the noise and detecting the presence of the wanted speech frequencies. In the case of dynamic background noise, particularly other speech, the problem of detecting the wanted speech signal becomes considerably harder. Any noise reduction system has to model the wanted speech signal and then try and remove, or at least attenuate, the other parts of the signal. Generally such a process will result in a speech signal which contains residual parts of the background sound where it overlaps with the wanted speech, and some parts of the speech signal will have been attenuated or suppressed where it was judged to be unwanted background. Whilst the result may be a cosmetically cleaner speech signal, the damage to the speech resulting from the processing has an impact on its intelligibility. Different techniques vary in the models they apply and the amount of attenuation applied [7] [8].

3.1 Speech Comprehension

Speech comprehension, especially in the presence of competing noise, is a predictive process based on incomplete information. The listener fills in the gaps in the information by trying to predict what the speaker is trying to say[1] [9]. This is obvious in the case of listening to someone with a stutter, where the listener is often tempted to fill in the word the speaker is having difficulty with. To achieve this level of prediction the listener uses their knowledge of the language, its structures and innate redundancy, their knowledge of the topic being talked about and any knowledge of the speaker and their normal speech patterns. The pace and rhythm, tone of voice and stress placement all form part of the information available to the listener. The listener may also lip-read and follow gestures to help support their comprehension. None of this supporting context is available to an audio processing algorithm, so any changes made to a speech signal may destroy more information than it reveals. Thus successful speech enhancement is a problem which involves modelling both the sound and the meaning of the speech. This is the same problem faced in the area of automatic speech recognition where the problem of competing background noise is still a major obstacle to success [10].

3.1 Audio Processing Algorithms and Intelligibility

The processing of speech to remove noise and improve intelligibility has been the subject of ongoing research over many years. Any situation where a speech signal is degraded with the addition of noise, reverberation or distortion is a candidate for the cleaning of the speech signal. It has, however, been recognised for many years that, whilst it may be possible to improve the perceived quality of a speech signal, improving the intelligibility of speech is far more difficult. Lim and Oppenheim, writing in 1979 point out that whilst many of the systems they examined could reduce the apparent background noise they often reduced intelligibility. They go on to point out that the separation of speech from background noise requires the modelling of the speech signal, and the better the model, the more successful the separation is likely to be. However, they also point out that the more assumptions the model makes about the speech the more likely it is to produces errors when these assumptions are invalid. Almost all the systems they tested reduced the intelligibility of the speech, and those that did not reduce intelligibility tended to degrade the quality of the signal [11].

A more recent review of a range of audio processing algorithms was published in 2007 by Philipos C. Loizou and his co-authors from the University of Texas. In two studies Hu and Loizou evaluated eight speech enhancement algorithms, for their impact on intelligibility, firstly with 25 subjects, a sentence recognition test, scoring the proportion of words identified correctly and in a follow-up test with 40 subjects testing for consonant recognition [12] [13]. They tested the speech with four types of noise at two different levels (noise at 0dB and 5dB below the speech level) against eight processed versions. They only found intelligibility improvements in one single noise condition (car interior 5dB below the speech) though the majority of the algorithms did not make the intelligibility any worse, the rest did reduce intelligibility. The algorithms that they had previously shown to perform well in terms of perceived quality [14] were not the same as the ones which performed the best in terms of intelligibility. They also found that different algorithms gave the best performance in different noise conditions, and that these differences were more apparent at the higher noise levels.

[1] This task can be understood as a case of the "inverse problem" where the listener is trying to create a model of the speaker's intentions from the sounds and gestures that the speaker makes.

4 Examples from Law Enforcement Research

In the field of law enforcement the problem of intelligibility is encountered when transcribing speech from field recordings. These are frequently of very poor quality due to the way in which they have been made. Whilst the people transcribing the recordings have normal hearing, the noise and distortion levels can be high enough to make comprehension of the material very difficult. Although the transcriber will often listen to a difficult passage several times to try and understand it, the parallels with TV viewing are useful, particularly as the listener is trying to understand the events and is listening for an hour or more at a time. Because this application is task based, it is possible to use performance based metrics, as well as single word identification intelligibility tests, to obtain objective measurements of the impact of signal processing. These tests move from a measure of narrow word intelligibility, to include the impact of issues of comfort, listening effort and fatigue on cognitive processing.

In a wide ranging paper presented at the 13th INTERPOL Forensic Science Symposium in 2001, A.P.A. Broeders comments that:

> *“Although the use of dedicated filtering hard- and software is widespread in the latter type of work, the net effect of the use of this equipment in terms of getting additional words down on paper is not always impressive. In fact, a large proportion of the work carried out under this heading is probably primarily of a cosmetic nature...” [15]*

He goes on to note that for people who have to transcribe large quantities of speech such filtering may increase productivity by reducing fatigue and recommends that for the best results in transcribing very low quality recordings the use of highly competent and educated native speakers of the language variety in use. He also notes that:

> *“A thorough familiarity with the accent and dialect of the speakers in the recording, as well as some familiarity with the details of the case, will often enable the analyst to compensate for the loss of redundancy of linguistic cues that is characteristic of poor quality recordings.”*

More recently in the UK the CLEAR project (Centre for Law Enforcement Audio Research) was set up in 2007 to target the needs of law enforcement in the UK for information about the latest technologies for speech cleaning. CLEAR is run jointly by Imperial College London (Department of Electrical and Electronic Engineering) and University College London (Department of Speech, Hearing and Phonetic Sciences) and funded by the UK Home Office [16].

The CLEAR project has used performance metrics to evaluate the impact of noise reduction. The project has taken the view that a mean opinion score of audio quality is insufficient to support the idea that noise reduction improves listening comfort and reduces fatigue.

The first test they used was named the “Typometer” [17]. It was a reaction time test requiring the subject to identify a spoken number in 5 different noise conditions. They tested good quality speech recordings against mixtures of the speech with car interior noise or babble noise, and the noise mixtures following noise reduction. Whilst the 20 subjects showed consistent levels of accuracy across the five conditions, their reaction times were significantly higher for the with-noise conditions than for the noise-free speech. The key finding was that the versions which had been processed using noise reduction were not significantly different from the with-noise versions [18].

The second test was dubbed the “Proofometer”. This used four-minute samples of spontaneous conversation in the same five noise conditions. Transcripts of the conversation with 50 errors, 30 word substitutions, 10 word deletions and 10 word insertions were created with care taken to disguise the errors so they could not be guessed from the transcript alone. The computer program displayed the transcript on screen and the subject was asked to click on the substituted or inserted word or the space where a word had been deleted. The error rates were then compared across the five noise conditions and the eighteen subjects. Again the versions with noise significantly degraded performance and there was no significant difference between the versions with and without noise reduction. A similar result was shown for the time taken for the subject to respond to the errors [19]. These two tests fail to support the hypothesis that noise reduction can lead to a reduced cognitive load, and thus an improvement in comprehension performance.

5 Broadcast and Film Industry Research

5.1 DICTION

DICTION - (Digitally Improving the Clarity of TelevisIOn Narrative for hard of hearing viewers) was an ITC project which ran from Feb 99 to Feb 2000, funded under the DTI/SERC Link programme [20]. The ambition was to produce a processor which could remove television background sound in the home, without degrading the main foreground sound, and thus enable the listener to hear more clearly [21] [22]. The processor was tested on a group of target users aged 63 to 84 years using the Revised Speech In Noise Test which asks the subjects to repeat the last word of each sentence they are played [23]. The test used 12-talker babble as the background noise and the final results were based on the results from 20 test subjects. The results showed that the algorithm was unsuccessful in meeting its aims of improving the intelligibility of speech for the target users [24].

5.2 AES Papers on Hearing Enhancement

There have been a number of more recent projects looking at similar issues for film as well as broadcast television. At the 2008 125th AES Convention in San Francisco there was a paper session given over to "Hearing Enhancement" at which a number of the following pieces of research were presented. This session was summarised by Francis Rumsey in a journal article the following year [25]. In that session two papers covered the issue of speech enhancement.

5.2.1 Dolby Laboratories

In a paper from the 125th AES Convention Hannes Müsch of Dolby gives a review of the literature on ageing and hearing loss and then discusses signal processing techniques which had the intention of generating "audio that is suited for elderly listeners" [26]. The processing was based around the assumption that 5.1 sound is in use and that the centre channel is carrying almost all of the speech signal in a programme. The processing attempted to attenuate non speech sounds, but only if they interfere with speech perception. The system used a speech detection system on the centre channel signal and multiband gain adjustment in the other channels based on the predicted intelligibility of the speech. However, the results of the processing described in the paper were not expected to increase word intelligibility, only reduce listening effort.

5.2.2 Fraunhofer Institute

Another paper from the same session describes work at the Fraunhofer Institute to reduce background sound in movies as part of the European project "Enhanced Digital Cinema" (EDCine, IST-038454) [27]. The paper describes an approach to enhancing the speech content and reducing non-speech components using speech detection and speech enhancement.. This work did not rely on any particular sound format, attempting to enhance speech that is already mixed into the same channel as other sounds. The listening tests were conducted on a group of 11 teenagers with impaired hearing and 12 people with normal hearing, a mixture of audio professionals and students. All subjects were asked to rate the results in terms of sound quality and speech quality on a scale of 1 to 10. The results for "sound quality", comparing the unprocessed sound to the proposed algorithm showed no change in perceived quality for the hearing-impaired group and a clear drop in quality for the normal hearing group. More importantly the results for "speech quality" showed no clear difference for the normal hearing group and a slight, but not statistically significant, drop in "speech quality" for the processed audio over the original.

5.3 DTV4All

DTV4All is a project funded by the European Commission, under the CIP ICT Policy Support Programme, to facilitate the provision of access services on digital television across the European Union [28]. It looked at a wide range of current and possible future access services. Amongst its shortlist of emerging access services it included improved audio for people with a hearing impairment, or “clean audio”. Whilst the project documents show that the original intention was to use clean dialogue from productions or from the centre channel of 5.1 mixes, [29] the resulting tests were conducted using pre-mixed soundtracks which had been “cleaned up”. The processing used in the “pre-tests” was carried out manually using the *Cedar 1500* processor. In the second set of pre-tests this was supplemented with an *Izotope* processor to remove “some impairing tonal components”. The main tests were carried out using the same processing combinations as the second set of pre-tests with the settings being varied manually for each test item [30] [31].

The test reports contain many comments from the tests about the difficulty of achieving good results from the audio processing for each of the different test items and the trade off between attenuation of background sounds and the tonal quality of the speech, particularly where there was spectral overlap with background music. The report gives insight into the quality of the processed speech, words like “tinny” and “sizzling” are used a number of times. The tests asked the participant to rate how well they could hear the dialogue and did not test for intelligibility. The results of the tests were very mixed, and are difficult to interpret. It is also worth noting that all the participants in the tests wore hearing aids so the processing of the sound by the hearing aids may have interacted with the processing being applied for the tests.

6 Conclusions

From the above examples it can be seen that current audio processing techniques cannot significantly improve the intelligibility of speech in noise, if at all. For the broadcaster this means that the production of “clean audio” from normal soundtracks through the use of noise reduction is not a viable option. The only way a clean speech signal can be created for the purpose of improving intelligibility is to capture clean, clear speech recordings at source.

More generally the message to programme production is that the intelligibility of speech recordings cannot be rescued by the use of noise reduction. Any impact on intelligibility from problems with speech recordings will continue to degrade the intelligibility of the speech even after noise reduction. Whilst such processing can be used to make the audio more pleasant on the ear, it cannot be used to restore the speech to its original level of intelligibility. This underlines the importance of good quality capture of speech in programme making as emphasised in the BBC’s new guidelines on TV sound [1].

7 References

[1] BBC, “Editorial Policy Guidance: Hearing Impaired Audiences”, March 2011. http://www.bbc.co.uk/guidelines/editorialguidelines/page/guidance-hearing-full

[2] Peter Olaf Looms, “The Case for DTV Access Services”, EBU Technical Review – 2010 Q2. http://tech.ebu.ch/docs/techreview/trev_2010-Q2_Access-Services1.pdf

[3] Danny Cohen, “Sound Matters”, BBC College of Production web site, March 2011. http://www.bbc.co.uk/academy/collegeofproduction/tv/sound_matters_cohen

[4] Voice of the Listener & Viewer, “VLV’s Audibility of Speech on Television Project will make a real difference”, VLV News Release 06/11, March 2011. http://www.vlv.org.uk/documents/06.11PressreleasefromVLV-AudibilityProject-0800hrs1532011_002.pdf

[5] “Clear Sound: best practice tips”, BBC College of Production web site, March 2011. http://www.bbc.co.uk/academy/collegeofproduction/tv/best_practice_tips

[6] James V. Stone, “Independent Component Analysis: A Tutorial Introduction”, The MIT Press, 2004.

[7] Yi Hu and Philipos C. Loizou, “A comparative intelligibility study of single-microphone noise reduction algorithms”, J.Acoust. Soc. Am. 122 (3), September 2007. http://www.utd.edu/~loizou/cimplants/intel_jasa07.pdf

[8] Yi Hu and Philipos C. Loizou, “A comparative intelligibility study of speech enhancement algorithms”, Proc. ICASSP, pages IV–561–564, Hawaii, 2007. http://ispl.korea.ac.kr/conference/ICASSP2007/pdfs/0400561.pdf

[9] Pichora-Fuller, M.K.,Schneider, B.A. and Daneman, M, “How young and old adults listen to and remember speech in noise”, J. Acoust. Soc Am. 97, 593-607, 1995.

[10] NIST STT Benchmark Test – May 09, http://itl.nist.gov/iad/mig/publications/ASRhistory/index.html

[11] Jae S. Lim & Alan V. Oppenheim, “Enhancement and Bandwidth Compression of Noisy Speech”, Proc. IEEE, Vol. 67, No. 12, December 1979. http://www.rle.mit.edu/dspg/documents/JLim1279.pdf

[12] Yi Hu & Philipos C. Loizou, “A Comparative Intelligibility Study of Speech Enhancement Algorithms”, Proceedings of ICASSP 2007. http://ispl.korea.ac.kr/conference/ICASSP2007/pdfs/0400561.pdf

[13] Yi Hu & Philipos C. Loizou, “A comparative intelligibility study of single-microphone noise reduction algorithms”, J. Acoust. Soc. Am. Vol. 122, issue 3, pp. 1777-1786 (Sept 2007). http://www.utdallas.edu/~loizou/cimplants/intel_jasa07.pdf

[14] Yi Hu & Philipos C. Loizou, “Subjective Comparison of Speech Enhancement Algorithms”, Proceedings of ICASSP 2006. http://www.utd.edu/~loizou/speech/subj_comparison.pdf

[15] A.P.A. Broeders, “FORENSIC SPEECH AND AUDIO ANALYSIS FORENSIC LINGUISTICS: 1998 to 2001 A Review, 13th INTERPOL Forensic Science Symposium, Lyon, France, October 16-19 2001. http://www.interpol.int/public/Forensic/IFSS/meeting13/Reviews/ForensicLinguistics.pdf

[16] Centre for Law Enforcement Audio Research web site, http://www.clear-labs.com/

[17] Huckvale, M., & Frasi, D., "Measuring the effect of noise reduction on listening effort", AES 39th International Conference on Audio Forensics, Copenhagen Denmark, June 2010. http://www.clear-labs.com/documents/aes2010pbmsq.pdf

[18] Huckvale, M., & Leak, J. "The effect of noise reduction on reaction time to speech in noise", Proc Interspeech 2009, Brighton, U.K., September 2009. http://www.clear-labs.com/documents/is2009-typometer-dist.pdf

[19] Huckvale, M., Hilkhuysen, G., & Frasi, D., "Performance-based Measurement of Speech Quality with an Audio Proof-reading Task", 3rd ISCA Workshop on Perceptual Quality of Systems, Dresden Germany, September 2010. http://www.clear-labs.com/documents/pqs2010proofometer.pdf

[20] DICTION project details collected on the Foundation for Assistive Technology web site, http://www.fastuk.org/research/projview.php?id=463

[21] Press Release: “Background noise suppression technology aids hard of hearing”, University of Surrey, August 2000. http://www.surrey.ac.uk/news/releases/8-1600bnst.html

[22] “Do not adjust your television set”, theEngineer, 14 August 2000 http://www.theengineer.co.uk/news/do-not-adjust-your-television-set/282075.article

[23] Bilger R., “Speech recognition test development”, ASHA Reports Number 14, American Speech-Language-Hearing Association, 1984. http://www.asha.org/uploadedFiles/publications/archive/ASHAReports14.pdf

[24] A.R. Carmichael, “Evaluating digital “on-line” background Noise suppression: Clarifying television dialogue for older, hard-of-hearing viewers”, Neuropsychological Rehabilitation: An International Journal, 1464-0694, Volume 14, Issue 1, 2004, Pages 241 – 249.

[25] Francis Rumsey, “Hearing Enhancement”, J.Audio Eng. Soc., Vol.57,No. 5, 2009 May.

[26] Hannes Müsch, “Aging and sound perception: Desirable characteristics of entertainment audio for the elderly”, 125th Convention of the Audio Engineering Society, paper 7627, October 2008.

[27] Christian Uhle, Oliver Hellmuth and Jan Weigel, “Speech enhancement of movie sound”, 125th Convention of the Audio Engineering Society, paper 7628, October 2008.

[28] DTV4All web site, http://www.psp-dtv4all.org/

[29] Werner Brückner & Ralf Neudel, “A Shortlist of Emerging Access Services”, DTV4All, Deliverable D3.1, December 2008. http://dea.brunel.ac.uk/dtv4all/ICT-PSP-224994-D31.pdf

[30] Werner Brückner, “Interim Report on Expert User Tests”, DTV4All, Deliverable D3.4, January 2010. http://dea.brunel.ac.uk/dtv4all/ICT-PSP-224994-D34.pdf

[31] “2nd Phase Emerging Access Service Demonstrators”, DTV4All, Deliverable D3.5, September 2010. http://dea.brunel.ac.uk/dtv4all/ICT-PSP-224994-D35.pdf