Towards Multi-modal Hearing Aid Design and Evaluation in Realistic Audio-Visual Settings: Challenges and Opportunities

Amir Hussain[1], *Jon Barker*[2], *Ricard Marxer*[2], *Ahsan Adeel*[1], *William Whitmer*[3], *Roger Watt*[1] *and Peter Derleth*[4]

[1]University of Stirling, UK, [2]University of Sheffield, UK,
[3]Unviersity of Nottingham, UK, [4]Sonova AG, Switzerland

[1]ahu@cs.stir.ac.uk [2]j.p.barker@sheffield.ac.uk

Abstract

A limited number of research developments in the field of speech enhancement have been implemented into commercially available hearing-aids. However, even sophisticated aids remain ineffective in environments where there is overwhelming noise present. Human performance in such situations is known to be dependent upon input from both the aural and visual senses that are then combined by sophisticated multi-level integration strategies. In this paper, we consider the opportunities and challenges presented by hearing-aid development in an audio-visual (AV) speech context. First, we posit the case for new multimodal AV algorithms that enhance speech quality and intelligibility with the aid of video input and low-latency combination of audio and visual speech information. Second, we consider the challenges that the AV setting presents to hearing aid evaluation. We argue that to meaningfully reflect everyday usage, hearing aid evaluation needs to be performed in an audio-visual setting regardless of whether hearing aids are directly using visual information themselves. We consider the need for new AV speech in noise listening tests, and for research into techniques for predicting objective AV speech quality and intelligibility. Finally, an AV speech enhancement evaluation challenge is proposed as a starting point for stakeholder discussion.

Index Terms: audio-visual speech, speech enhancement, speech intelligibility assessment

1. Introduction

The multimodal nature of speech is well established. Speech is produced by the vibration of the vocal cords being filtered according to the configuration of articulatory organs. Due to the visibility of some of these articulators (i.e., lips, teeth and tongue), there is an inherent and perceptible relationship between audio and visual speech properties. Pioneering work [1, 2, 3] demonstrated that listeners exploit this relationship, unconsciously lip reading to improve the intelligibility of speech in noise [4]. Further, looking at a speaker makes speech more detectable in noise [5], i.e., as if audio cues are being visually enhanced [6].

Embracing the multimodal nature of speech presents both opportunities and challenges for hearing assitive technology: on the one hand there are opportunities for the design of new multimodal algoirthms; on the other hand multimodality challenges the current standards for hearing aid evaluation, which generally consider the perception of the audio signal in insolation.

This paper will first consider the potential benefits of designing fully audio-visual hearing devices. In particular, we consider the design of a new breed of device that employs both microphones and video sensors. Such a device has the potential to extract information from the pattern of the speaker's face and lip movements and to employ this information as an additional input to speech enhancement algorithms. In Section 2 we discuss the AV-COGHEAR project that is aiming to build and test prototypes of this technology.

In Section 3 we turn attention to the challenge of hearing device evaluation. Our main concern in this regard is that standard evaluation strategies, which use an audio-only setting, may not be predictive of a device's performance when used in real multimodal conditions. This is just as much true for devices that use audio-only input as it is for audio-visual devices. We consider the requirements of a fully multimodal evaluation and conclude in Section 4 by making a proposal for an open multimodal speech enhancement challenge that we hope will stimulate fresh research in this area.

2. Audio-Visual speech enhancement

2.1. Background

Despite decades of research, there are are few speech enhancement algorithms that can reliably increase the intelligibility of speech corrupted by complex noises typical of everyday listening conditions. For example, spectral subtraction can be very effective for reducing the perception of noise in stationary conditions, but the apparently 'cleaner' processed speech turns out to be no easier to understand. If multiple microphones are available then beamforming algorithms can lead to genuine speech intelligibility improvements but even these techniques are hard to employ in an unpredictable noise environment. Consequently hearing aid algorithms achieve most of their benefit simply by amplifying the signal into the audible range, and offer little advantage for speech listening when speech is present in high levels of background noise.

There is reason to believe that, in contrast to audio-ony algorithms, audio-visual speech enhancement approaches may be able to offer consistent intelligibility gains – especially for hearing impaired listeners. To understand why visual features may be beneficial it is important to understanding why noise renders speech less intelligible in the first place. The commonly understood view is that the noise sources reduce speech intelligibility by energetically masking the target source. Visual signals can then restore intelligibility by delivering phonetic information that has been obliterated in the masked regions. However, this is only part of the picture. Intelligibility is also governed by *informational* masking (IM), i.e., the degree to which the auditory system is able to, i) segregate spectro-temporal (ST) regions that are speech dominated from those that are background-dominated, and ii) focus attention on the target regions.

IM is amplified by even mild hearing impairment, leading to large speech intelligibility losses in social situations where speech is present in the noise background. It has been suggested that this is partly due to the loss of precision with which ‘grouping cues’ are encoded - i.e., signal properties such as periodicity and apparent location that allow for a signal to be sequentially organised, [7]. Schwartz et al. [8] have shown that visual cues can be exploited at a pre-phonetic stage, reducing IM. Essentially, visual cues can supplement auditory grouping cues, providing a signal that directs attention to the ST regions dominated by the target source.

We believe that listeners may benefit from an AV hearing device that is able to mimic the IM releasing function of visual cues. For example, the device would use the visual information to direct the audio signal processing to amplify speech signal components and attenuate noise components.

2.2. The AV-COGHEAR project

The ongoing UK Engineering and Physical Sciences Research Council (EPSRC) funded AV-COGHEAR project, collaboratively led by Stirling and Sheffield Universities, is a first attempt at developing a cognitively-inspired, adaptive and context-aware approach for combining audio and visual cues (e.g., from lip movement) to deliver speech intelligibility enhancement [9].

The project’s overarching goal is the development of next-generation multi-modal hearing aids and listening devices that have the potential to be a disruptive technology redefining user expectations. Beyond hearing aid devices we foresee impact in a number of areas including: cochlear implant signal processing, speech recognition systems, auditory systems engineering in general, and clinical, computational, cognitive and auditory neuroscience. A preliminary deep-learning-driven, multi-modal speech enhancement framework pioneered at Stirling [10] is currently being significantly extended to incorporate innovative perceptually-inspired models of auditory and AV scene analysis developed at Sheffield [11]. Further, novel computational models and theories of human vision developed at Stirling are being deployed to enable real-time tracking of facial features. Contextual multimodality selection mechanisms are being explored, and collaborations with SONOVA and MRC IHR, will facilitate envisaged delivery of a clinically-tested software prototype.

In the literature, much progress has been made to develop enhanced speech processing algorithms capable of improving speech quality. In contrast, little work has been conducted to design algorithms that can improve speech intelligibility. In this project, our hypothesis is that it is possible to combine visual and acoustic input to produce a multimodal hearing device that is able to significantly boost speech intelligibility in the everyday listening environments, in which traditional audio-only hearing devices prove ineffective.

To test this hypothesis, we are collaboratively working to develop and clinically validate a next-generation cognitively-inspired, AV hearing-device software prototype, capable of real-time implementation, which will autonomously adapt to the nature and quality of its visual and acoustic environmental inputs. In this context, we have currently developed two contrasting approaches to speech enhancement developed respectively at Stirling and Sheffield: (1) A deep lip-reading driven Weiner filtering approach, shown in Figure 1 and (2) an audio-visual analysis-resynthesis approach, depicted in Figure 2 [12]. The preliminary objective and subjective evaluation has revealed the potential and reliability of the proposed AV technology as compared to the state-of-the-art audio only speech processing techniques.

3. Speech enhancement evaluation in realistic AV settings

3.1. Background

The development of hearing devices that utilise both audio and visual information, highlights the growing need for hearing devices to be evaluated in realistic multimodal settings. In the literature, there exist several standards for evaluating hearing aid algorithms in audio only settings, ranging from the Connected Speech Test (CST) (CST; [13]), the Speech Intelligibility Index (SSI) [14], to Kates’ extension to the SSI [15]. However, there are no established standards for evaluating hearing-aid algorithms in audio-visual settings. Note, an audio-visual extension of the CST [16] was proposed shortly after the audio-only test but has not been widely adopted.

Evaluating a hearing device in an audio-only setting may produce a misleading view of the speech intelligibility benefits it will provide. Except in a few naturally audio-only situations, (e.g., telephone conversations), hearing aid users who are struggling to understand speech in noise will be closely attending the speaker’s lips. These listeners will therefore experience a visual benefit which will improve their aided-performance. Note, this is true regardless of whether the hearing aid is using the visual signal itself. The size of this visual benefit needs to be accounted for.

Visual-benefit would not be a problem for hearing aid evaluation if the size of the benefit was independent of the hearing aid algorithm. If this was true then the ranking of algorithms would remain the same and a best algorithm could still be chosen. However, this is unlikely to be the case. For example, consider an algorithm that emphasises aspects of the acoustic signal that are redundantly encoded in the visual signal. This algorithm might provide large benefits in an audio-only evaluation but then be shown to proffer little benefit in a setting where the user sees the lip movements directly.

Despite the clear necessity for hearing aid speech performance to be evaluated from an AV perspective, there has been very little work in this direction. Recently, Wu and Bentlier [17] examined how visual cues impact directional benefit and preference for the directional microphone hearing-aid. The authors administered two speech recognition in noise tests to assess directional benefit: (1) the AV version of the Connected Speech Test (CST; [13, 16]) and (2) the Hearing in Noise Test [18] to investigate the impact of visual cues on the directional benefit. It was reported that visual cues significantly improved speech recognition performance to its ceiling level and reduced the directional benefit and preference for directional processing.

3.2. Challenges for audio-visual hearing aid evaluation

In this section we outline the main challenges facing audio-visual speech intelligibility testing.

3.2.1. Audio-visual HA performance predictors

In an ideal world, hearing aid algorithms could be evaluated cheaply using algorithms that would predict the intelligibility and/or quality of the processed speech. The processed audio speech signal and its video counterpart could be fed into an objective test that would accurately predict intelligibility of the signal. Unfortunately, this is an unrealistic proposition which

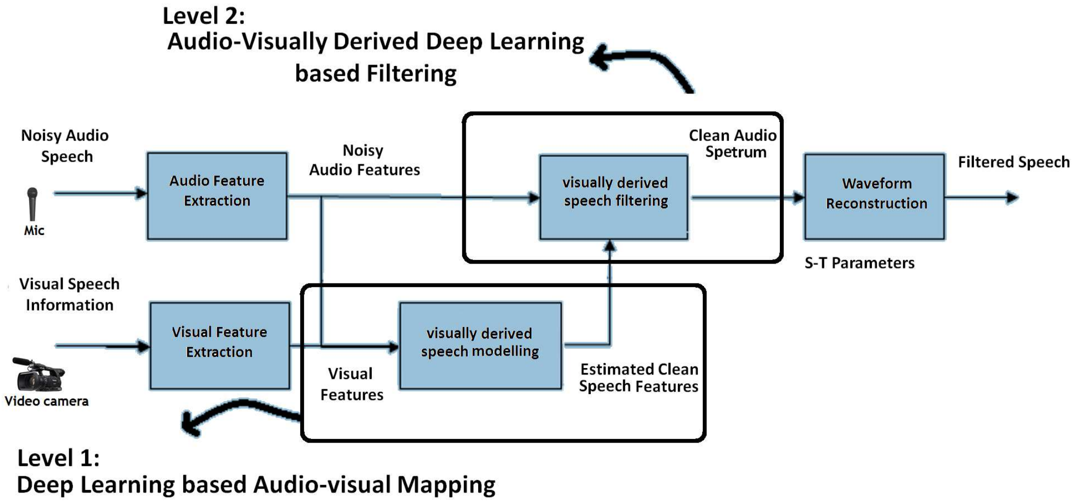
Level 2:
Audio-Visually Derived Deep Learning based Filtering
Noisy Audio Speech
Mic
Audio Feature Extraction
Noisy Audio Features
visually derived speech filtering
Clean Audio Spetrum
Waveform Reconstruction
Filtered Speech
S-T Parameters
Visual Speech Information
Video camera
Visual Feature Extraction
Visual Features
visually derived speech modelling
Estimated Clean Speech Features
Level 1:
Deep Learning based Audio-visual Mapping

Figure 1: *Deep lip-reading driven Weiner filtering (Stirling)*

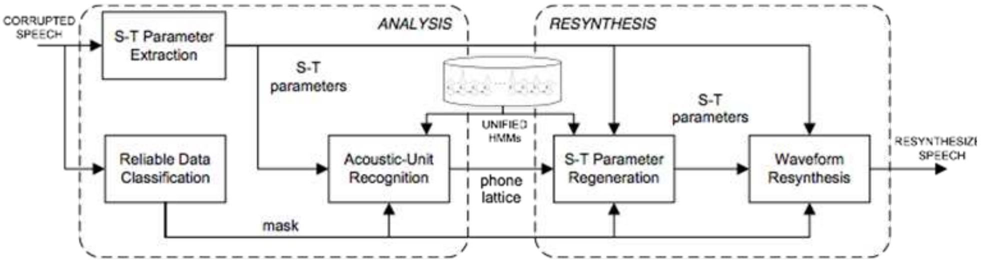

Figure 2: *Audio-visual analysis-resynthesis framework (Sheffield)*

remains a challenge even for audio-only evaluation.

There have been many proposed metrics for objective speech quality and intelligibility prediction. Algorithms are categorised as intrusive or non-intrusive depending on whether they require a clean speech reference signal or not, respectively. We can assume that for hearing-aid development a reference signal can be available and therefore intrusive algorithms can be applied. These include the normalized covariance metric (NCM) [19] and short-time object intelligibility (STOI) [20] which predict intelligibility and perceptual evaluation of speech quality (PESQ) [21] which predicts speech quality. Although different in detail, these algorithms all operate by making a weighted comparison between an auditorily-inspired representation of the reference and corrupted signal.

More recently developed predictors have been especially designed for hearing aid (HA) processing. These include the HA speech quality index (HASQI) [22], the HA speech intelligibility index (HASPI) [23], and an extension of the perception-model-based quality prediction method (PEMO-Q) [24] adapted for hearing impairment (PEMO-Q-HI) [25]. These predictors again compare a reference and a processed signal in an auditory model space, however, their auditory models can be tuned to mimic the effects of a listener's hearing impairment, (e.g., raised thresholds, filter broadening, etc).

A particular problem with these approaches is that their performance can be sensitive to the type of processing performed by the aid (see [26] for a review). For example, non-linear frequency compression (NFC) – a recent development in hearing aids which warps the signal spectrum to fit the listener's usable frequency range – can generate big apparent differences between the reference and processed signal. Unless the metric is designed to expect NFC and compensate for this frequency warping it will predict the NFC processed signal to have low quality/intelligibility. The fundamental problem here is that the metrics are necessarily built on shallow models of speech perception. The resulting need to fit the prediction models to hearing aid algorithms is surely problematic if they are expected to effectively evaluate novel and unanticipated approaches to hearing aid signal processing.

When considering audio-visual intelligibility the situation is worse. There are no adequate models of how acoustic and visual speech information are combined during speech perception. There are currently no models that can predict effects such as informational masking release in non-stationary masker conditions. Without this understanding there is no basis on which to start building models of AV speech intelligibility.

3.2.2. *Difficulties with speech-in-noise listening tests*

There are a large number of standard speech-in-noise listening tests that can be used to directly measure the intelligibility benefit of a hearing device. They include adaptive and fixed SNR tests. The former include the Hearing in Noise Test

(HINT) [27], QuickSIN [28], Words in Noise (WIN) [29, 30] and Bamford-Kowal-Bench SIN (BKB-SIN) [31, 32]. These tests automatically adapt the SNR to the threshold at which communication breaks down (i.e., at which a fixed percentage of words are incorrectly recognised). They are quick and easy to administer but have the drawback of not being able to provide information about performance at other SNRs above and below this threshold. Fixed SNR tests such as CST [13] and Speech Perception in Noise Test (SPIN) [33] measure the percentage of words correctly recognised at a presentation SNR. However, they are susceptible to ceiling effects, i.e., no benefit can be measured once all words are recognised correctly. Choosing an appropriate SNR can be difficult.

Tests vary with respect to the type of speech material - some using isolated words and others complete sentences. Sentences are regarded as more appropriate materials for intelligibility measurement as they reflect real speech and produce steeper psychometric functions that more accurately estimate threshold SNRs. Sentences need to be carefully designed to be phonetically balanced and to have low predictability. So-called 'matrix tests' achieve this by using randomly chosen words from the closed-set in each word position in a sentence: e.g., a sentence might be composed as: <name>, <verb>, <number>, <adjective>, <noun> with 10 choices for each slot leading to 100,000 possible sentences. Such tests have been designed for many different languages, e.g. German [34, 35], Spanish [36], English [37], etc. There has been recent success in predicting sentence test performance in a wide range of noise conditions using statistical techniques adapted from the speech recognition community [38].

Speech-in-noise tests have primarily been designed for use in clinical settings for fitting a device to a given user. Comparing two devices for a given user is a well-posed problem. However, if a conclusion is required about which device is better in a more general sense, then obvious problems emerge. The question is impossible to answer independently of some characterisation of the user, i.e., the precise nature and degree of their hearing deficit. The test would then require a pool of listeners matching this characterisation that is large enough to average out remaining individual differences. This is particularly problematic given that audiograms – the standard characterisation of hearing deficit – can be poor predictors of speech in noise performance (and even poorer predictors of audio-visual speech recognition ability).

The difficulties experienced with audio-only testing are compounded in audio-visual settings, specifically: the visual cues may make it more likely to encounter ceiling effects; the test-retest scores are likely to be more variable given the increased cognitive complexity of the task; there is a starker contrast between controlled, well-framed, clearly articulated visual input and video characteristic of everyday conversational settings; and selecting homogeneous listener pools is more challenging as there are large and unpredictable individual differences in visual speech benefit amongst listeners.

3.2.3. The need for a realistic AV corpus

Reliable evaluation of future AV speech filtering technology will require subjective intelligibilty assessment. This raises the question of what type of speech material to use. Although there exist a number of small well-controlled audio-visual speech corpora, such as BANCA [39], AVICAR [40], VidTIMIT [41], and Grid [42], there is a need for evaluation of multi-modal speech enhancement systems using realistic audiovisual speech data. Audio-visual datasets are required in which speakers are speaking more naturally than in many existing corpora, including conversational speech and imperfect visual data. This is represented by the speaker moving their head, obscuring their face, and also different levels of background noise to take account of the Lombard effect (where speakers naturally adjust their speech to take account of different levels of background noise). To our knowledge, there is no corpus available that contains a sufficient range of AV speech data, or variety of A and V noise (i.e., acoustic noise, speaker movement and occlusion, etc.).

4. Conclusion – Towards an open AV evaluation framework

Future multimodal hearing devices will demand new approaches to evaluation. There will be a need to reconsider how hearing aid algorithms are evaluated during development so as to incorporate visual input, subject to real-time, low-latency constraints. There is also need to reconsider how devices are prescribed and fitted to patients. It will no longer be appropriate to use audio-only speech-in-noise tests. For the new devices the key question will be how the AV processing of the device interacts with the AV processing of the user when presented with realistic AV input.

New standards will only emerge from discussion across the sector involving manufacturers, health care professionals, audiologists and patients. We hope the 2017 CHAT Workshop will provide a starting point for this activity. To stimulate progress we plan to organise an open competition for AV speech enhancement evaluation, which could be run as part of an international INTERSPEECH Workshop in 2018. The competition could run along similar lines to the ISA Blizzard Challenge where the cost of participation pays for evaluation of the algorithms.

We conclude by presenting a tentative proposal for hearing algorithm evaluation with the aim of seeking feedback from the community.

The evaluation campaign will have two phases:

Phase 1 - development of enhancement algorithms: The proposal would be to use the existing AV Grid corpus [42] as the source material. This has a design similar to a matrix test and has already had extensive use in audio-only speech intelligibility studies. We would then mix this corpus with everyday complex noise backgrounds such as café and street noises recorded in the CHiME-3 corpus [43]. Participants would be invited to apply their algorithms (audio-only or audio-visual). We would then measure subjective intelligibility and subjective quality using a large bank of paid listening subjects who would be presented with the processed audio alongside the video. The evaluators could be NH or a homogenous group of HI listeners.

Phase 2 - development of new objective measures: Phase 1 will generate a large amount of listener data, i.e., showing how listeners have responded to variously-processed noisy AV speech samples. This data can then be used to test models that predict AV intelligibility and speech quality. We could immediately evaluate existing audio-based predictors which would be expected to underestimate AV performance. The challenge would then be to develop new AV predictors that extend these models. We would release a subset of the data from Phase 1 for model development and retain a hidden set for model evaluation.

5. Acknowledgements

This research is funded by the EPSRC project AV-COGHEAR, EP/M026981/1.

6. References

[1] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *The journal of the acoustical society of america*, vol. 26, no. 2, pp. 212–215, 1954.

[2] N. P. Erber, "Auditory-visual perception of speech," *Journal of Speech and Hearing Disorders*, vol. 40, no. 4, pp. 481–492, 1975.

[3] Q. Summerfield, "Use of visual information for phonetic perception," *Phonetica*, vol. 36, no. 4-5, pp. 314–331, 1979.

[4] E. Vatikiotis-Bateson, I.-M. Eigsti, S. Yano, and K. G. Munhall, "Eye movement of perceivers during audiovisualspeech perception," *Attention, Perception, & Psychophysics*, vol. 60, no. 6, pp. 926–940, 1998.

[5] J. Kim and C. Davis, "Investigating the audio–visual speech detection advantage," *Speech Communication*, vol. 44, no. 1, pp. 19–30, 2004.

[6] K. W. Grant and P.-F. Seitz, "The use of visible speech cues for improving auditory detection of spoken sentences," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1197–1208, 2000.

[7] D. Ruggles, H. Bharadwaj, and B. G. Shinn-Cunningham, "Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication," *Proceedings of the National Academy of Sciences*, vol. 108, no. 37, pp. 15 516–15 521, 2011.

[8] J.-L. Schwartz, F. Berthommier, and C. Savariaux, "Seeing to hear better: evidence for early audio-visual interactions in speech identification," *Cognition*, vol. 93, no. 2, pp. B69–B78, 2004.

[9] A. Abel, R. Marxer, J. Barker, R. Watt, W. Whitmer, P. Derleth, and A. Hussain, "A data driven approach to audiovisual speech mapping," in *Advances in Brain Inspired Cognitive Systems. BICS 2016*, C. Liu, A. Hussain, B. Luo, K. Tan, Y. Zeng, and Z. Zhang, Eds. Spring, 2015, vol. 10023, pp. 331–342.

[10] A. Abel and A. Hussain, "Novel two-stage audiovisual speech filtering in noisy environments," *Cognitive Computation*, vol. 6, no. 2, pp. 200–17, 2014.

[11] J. Barker and X. Shao, "Energetic and informational masking effects in an audio-visual speech recognition system." *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 3, pp. 446–458, Mar. 2009.

[12] J. Carmona, B. J., A. Gomez, and N. Ma, "Speech spectral envelope enhancement by HMM-based analysis/resynthesis," *IEEE Signal Processing Letters*, vol. 20, no. 6, pp. 563–66, 2013.

[13] R. Cox, G. Alexander, and C. Gilmore, "Development of the Connected Speech Test," *Ear and Hearing*, vol. 8, no. 5, pp. 119S–126S, 1987.

[14] ANSI S3.5-1997, "American National Standard methods for the calculation of the Speech Intelligibility Index," American National Standards Institute, New York, Tech. Rep., 1997.

[15] J. Kates, "Coherence and the Speech Intelligibility Index," *The Journal of the Acoustical Society of America*, vol. 115, no. 2604, pp. 1085–99, 2004.

[16] R. Cox, G. Alexander, C. Gilmore, and P. K.M., "The Connected Speech Test version 3: Audiovisual administration," *Ear and Hearing*, vol. 10, no. 1, pp. 29–332, 1989.

[17] Y. Wu and R. Bentlier, "Impact of visual cues on directional benefit and preference: Part I–laboratory tests." *Ear and Hearing*, vol. 31, no. 1, pp. 22–34, 2010.

[18] M. Nilsson, S. Soli, and J. Sulivan, "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–99, 1994.

[19] I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *The Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1703–1716, 1996.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[21] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)–a new method for speech quality assessment of telephone networks and codecs," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 749–752.

[22] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI)," *Journal of the Audio Engineering Society*, vol. 58, no. 5, pp. 363–381, 2010.

[23] ——, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[24] R. Huber and B. Kollmeier, "PEMO-QA new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 6, pp. 1902–1911, 2006.

[25] R. Huber, V. Parsa, and S. Scollie, "Predicting the perceived sound quality of frequency-compressed speech," *PloS one*, vol. 9, no. 11, p. e110260, 2014.

[26] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE signal processing magazine*, vol. 32, no. 2, pp. 114–124, 2015.

[27] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–1099, 1994.

[28] M. C. Killion, P. A. Niquette, G. I. Gudmundsen, L. J. Revit, and S. Banerjee, "Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2395–2405, 2004.

[29] R. H. Wilson, "Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance," *Journal of the American Academy of Audiology*, vol. 14, no. 9, pp. 453–470, 2003.

[30] R. H. Wilson and C. A. Burks, "Use of 35 words for evaluation of hearing loss in signal-to-babble ratio: A clinic protocol," *Journal of rehabilitation research and development*, vol. 42, no. 6, p. 839, 2005.

[31] P. Niquette, J. Arcaroli, L. Revit, A. Parkinson, S. Staller, M. Skinner, and M. Killion, "Development of the BKB-SIN test," in *Proc. annual meeting of the American Auditory Society*, 2003.

[32] J. Bench, A. Kowal, and J. Bamform, "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," *British Journal of Audiology*, vol. 13, pp. 108–112, 1987.

[33] D. N. Kalikow, K. N. Stevens, and L. L. Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *The Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1337–1351, 1977.

[34] K. Wagener, T. Brand, V. Kühnel, , and B. Kollmeier, "Entwicklung und evaluation eines satztests fürdie Deutsche sprache I: Design des Oldenburger Satztests (Development and evaluation of a sentence test for the German language I: Design of the Oldenburg Sentence Test)," *Z. Fuer Audiologie, Audiological Acoust*, vol. 38, no. 2, pp. 4–15, 1999.

[35] ——, "Entwicklung und evaluation eines satztests fürdie Deutsche sprache II: Optimierung des Oldenburger Satztests (development and evaluation of a sentence test for the German language II: Optimization of the Oldenburg Sentence Test)," *Z. Fuer Audiologie, Audiological Acoust*, vol. 38, no. 2, pp. 44–56, 1999.

[36] S. Hochmuth, T. Brand, M. Zokol, F. Zenker Castro, N. Wardenga, and B. Kollmeier, “A Spanish matrix sentence test for assessing speech reception thresholds in noise.” *Int. J. Audiol.*, vol. 51, no. 7, pp. 536–544, 2012.

[37] M. Zokoll, A. Warzybok, R. Carroll, B. Kreisman, P. Allen, K. Wagener, and B. Kollmeier, “Design, optimization, and evaluation of an American English matrix sentence test in noise.”

[38] B. Kollmeier, M. R. Schädler, A. Warzybok, B. T. Meyer, and T. Brand, “Sentence recognition prediction for hearing-impaired listeners in stationary and fluctuation noise with FADE: Empowering the attenuation and distortion concept by plomp with a quantitative processing model,” *Trends in Hearing*, vol. 20, pp. 1–17, 2016.

[39] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée *et al.*, “The BANCA database and evaluation protocol,” in *International conference on Audio-and video-based biometric person authentication*. Springer, 2003, pp. 625–638.

[40] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. S. Huang, “AVICAR: audio-visual speech corpus in a car environment.” in *INTERSPEECH*, 2004, pp. 2489–2492.

[41] C. Sanderson, “The VidTIMIT database,” IDIAP, Tech. Rep., 2002.

[42] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[43] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third chimespeech separation and recognition challenge: Dataset, task and baselines,” in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 504–511.