Efficient intelligibility evaluation using keyword spotting

A study on audio-visual speech enhancement

EFFICIENT INTELLIGIBILITY EVALUATION USING KEYWORD SPOTTING: A STUDY ON AUDIO-VISUAL SPEECH ENHANCEMENT

Cassia Valentini-Botinhao, Andrea Lorena Aldana Blanco, Ondrej Klejch, Peter Bell

The Centre for Speech Technology Research, University of Edinburgh, UK

ABSTRACT

We propose a new method for human speech intelligibility evaluation based on keyword spotting. In this method, participants play a stimulus and select the word they hear from a close set of alternatives. To find which sentence to use, the target word, and alternatives we mine a large set of stimuli using a phonetic dictionary and a language model. Unlike other tests, our method does not rely on specially designed sentences and can be used to evaluate in-the-wild material such as TED talks. We focus on audio-visual (AV) speech enhancement (SE) evaluation as a study case. We compared our method to a transcription task and observed that the two produce highly correlated results, albeit our task requiring substantially less participation time. We then adopted it on a large-scale evaluation of AVSE systems. Results show that keyword spotting is a suitable and efficient alternative to assess intelligibility from AV stimuli.

Index Terms— Intelligibility evaluation, keyword spotting, audio-visual speech enhancement

1. INTRODUCTION

Intelligibility evaluation is crucial in the design and development of speech enhancement or synthesis technologies to be consumed by humans, for example, in hearing aids, personal assistants, public announcement systems, navigation systems and assistive communication devices. Beyond the demand of quality and naturalness, users of such technologies benefit from higher levels of intelligibility, particularly in adverse scenarios that hinder effective communication. A system's intelligibility can be evaluated using objective or subjective methods. The main advantage of objective evaluation is that it requires fewer resources: it is faster and cheaper to carry out. However, objective metrics such as the short-time objective intelligibility (STOI) [1] have shown to not necessarily be good predictors of human performance [2, 3]. Additionally, even though several metrics to predict intelligibility have been proposed [3, 4] and each measure operates in a deterministic manner, particular characteristics of the stimuli – including characteristics of the speaker and listener, the linguistic content, the noise, the quality of the recording – can lead to drastically different results. Overall, designing a measure that works well across different stimuli conditions is a very challenging task.

A common alternative approach is to use "subjective" evaluation methods in which human participants are asked to listen to a set of stimuli and provide some sort of feedback. The most common type of intelligibility evaluation is based on a transcription task, where participants are asked to transcribe what they heard. In this paradigm the intelligibility of a sentence is scored by comparing the spoken text to the text transcribed by participants. In most cases this comparison is in the form of word accuracy calculation (i.e., the percentage of words correctly transcribed) [5]. Transcription tasks tend to use carefully designed sentences that are phonetically balanced and where word predictability is to some extent controlled or manipulated. In some applications however it is not possible to record new material for evaluation and we need to make use of existing stimuli. Adopting a transcription task – and the accompanying scoring method – on sentences that are not specifically designed for this task is relatively inefficient. Asking participants to transcribe a complete sentence when in reality only a few set of words is highly unpredictable makes it a very time consuming task, whilst the differing ability of participants to leverage prior knowledge in completing the transcription task may make results hard to compare across stimuli. Moreover, selected evaluation material should reflect a more realistic scenario that users of the technology are going to experience.

In our specific use case of AV speech enhancement (AVSE) evaluation, there have been attempts to create AV datasets for subjective testing of AV stimuli, but they have been used to evaluate attributes different to intelligibility (i.e., audio-visual benefit [6] and quality [7]); and video material is normally collected using a fixed shot in which the speaker is always seen in the same position. Although a range of AVSE systems have been proposed [8, 9, 10, 11, 12], the area is relatively new: systems are mainly evaluated using objective metrics and when subjective evaluation takes place is conducted using audio-only stimuli [7, 13]. This mismatch in the evaluation method and the real-world application makes it difficult to properly assess the contribution of the audio-visual speech enhancement systems. It is then crucial to develop suitable AV subjective evaluation protocols and evaluation stimuli that allow reliable performance assessment of audio-visual SE systems.

In this paper we present a new method for evaluating intelligibility based on keyword spotting. We present participants with AV stimuli that comprise video and audio of a target speaker mixed with an interferer. Participants have to select the word that they heard out of a closed set of alternatives. To design this closed set, choose target words and sentences we mine pre-existing stimuli using a phonetic dictionary and a language model. To validate the method we present details of the first wide scale evaluation of audio-visual speech enhancement systems, performed for the AVSE 2022 Challenge [14]. We show

how the proposed paradigm resembles a transcription-based task in terms of ranking intelligibility of stimuli mixed with different maskers at different levels and that the new method is suitable for ranking the performance of AVSE systems,

2. BACKGROUND: INTELLIGIBILITY EVALUATION

There are a number of protocols available for evaluating intelligibility with human listening. Lexical decision tasks involve identifying whether the sample heard is a dictionary word or a nonword [15]. Word recognition tasks require the participant to identify which word was spoken or at which point a word becomes identifiable. Word recall tasks involve asking participants which words or sentence they heard, testing for instance the effect on memory and cognitive overload. The most common type of test used to evaluate SE systems is the transcription test, which requires the participant to type in what they hear. The results can then be compared to the text used to generate the stimuli to provide a measure of the intelligibility through word accuracy rates or phoneme error rates. A design decision crucial to transcription tasks is the choice of sentences or words. Some of the specially designed sentence material used in intelligibility evaluation are the matrix sentences [16], the semantically unpredictable sentences (SUS) [17], the Harvard sentences [18] and the SPIN sentences [19]. They differ in terms of syntactic structure, varying from fixed structure (matrix sentences) to a more flexible structure (Harvard), and in terms of predictability. SUS are made of words that are not predictable by context; matrix sentences are made by a closed set of words; the last word in the SPIN sentences is either of low or high predictability. They also differ in respect of whether every word is scored equally (matrix, SUS, Harvard) or not (SPIN); and whether participants are presented with a closed set of responses (matrix) or not.

The choice of sentence material has a big impact on the task: closed set tests are easier to perform than open set tests; highly unpredictable material is harder than structured material; transcription tests take longer, require more from participants, and demand some kind of post processing to correct for typos. Moreover, when evaluating the intelligibility of speech in noise, the sentence material will have a big influence on the levels of noise that participants should be exposed: easier material will require higher noise levels to reach the same performance, making the task less realistic.

3. KEYWORD SPOTTING EVALUATION

We propose a new method for intelligibility evaluation that can be used to evaluate in-the-wild stimuli (reflecting more realistic material) while sharing some of the advantages of specially designed closed set tasks. Our method is based on keyword spotting.

In our approach human participants are presented with a video clip of a spoken utterance of around 5-10s in length. Participants are asked to select one word from a list of alternatives according to what they heard in the video clip. Specifically, participants are asked the question "which of the following words did the speaker in the video say?" In our case the list of options contains four alternatives and a special alternative "none of the above". We ensure that only one of the alternatives was actually spoken, but also set the "none of the above" alternative to be just as likely to be correct as the word options – so in 20% of stimuli we do not include any word spoken in the utterance. Figure 1 presents an example of the chosen sentence and the alternatives (including the target keyword). In an ideal situation we would record the test stimuli using predefined utterances to have a very fine control over the lexical content. However, this is not possible in situations where it is necessary to make use of pre-existing speech data. Next, we describe the data mining process to implement keyword spotting in an evaluation task from an already existing set of recordings and their transcripts.

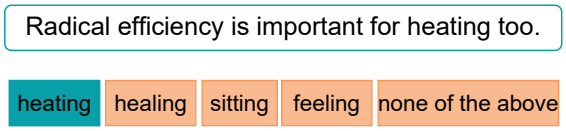

Fig. 1: Keyword spotting example (top: selected sentence; bottom: keyword and alternatives). The keyword (target word) is shown in blue.

Our data mining procedure can be split into three parts: finding phonetically similar words for each word, identifying target words in each sentence, and selecting sentences for evaluation. In the first part we create a list of similar sounding words for each word in the transcripts. We find the alternatives by using a phonetic dictionary and computing edit distance between all pairs of words. Instead of using the classic edit distance algorithm we compute substitution cost as a distance between articulatory features of the two phones in question. We select an appropriate threshold for the maximum phonetic similarity distance between two words to accept them as phonetically similar.

Once we have a set of alternatives for each word in the sentence, the next step is to select a word in the sentence which should be used for the keyword spotting task. Since we do not have control over the lexical content of the sentence, we wish to find hard words as evaluation targets. Moreover, we want to use words whose alternatives are meaningful substitutions for the target words, so as not to make the evaluation task too simple. We can achieve this by using a statistical language model trained on in-domain data. This can be formalized as follows. Let the sequence of words be denoted as $Y = (y_1, ..., y_n)$, we define a function $f(Y, i, a)$, that replaces the ith word in the word sequence Y with an alternative a, such that:

$$f(Y, i, a) = (y_1, ..., y_{i-1}, a, y_{i+1}, ..., y_n). \quad (1)$$

Furthermore, we assume that we have access to a n-gram language model $P(X)$, which can be factorised as:

$$P(Y) = \prod_{i=1}^{n} P(y_i \mid y_1 ... y_{i-1}) \quad (2)$$

We can use this language model to compute perplexity of a sequence of words as:

$$ppl(Y) = exp\{-\frac{1}{n}\sum_{i=1}^{n} \log P(y_i \mid y_1 ... y_{i-1})\}. \quad (3)$$

Let us also denote function $g(w)$, which for each word w returns a set of phonetically similar words. For each sentence Y we find the target word i as the word whose worst alternative (i.e. the alternative that maximises perplexity) achieves the best (i.e the lowest) perplexity across the worst alternatives of all words in the sentence:

$$i = \arg \min_{i \in \{1..n\}} \max_{a \in g(y_i)} ppl(f(Y, i, a)). \quad (4)$$

In the last step of the data mining pipeline, we aim to select a subset of sentences which should be presented to human listeners. Despite our efforts to find the most phonetically similar and most meaningful alternatives for each sentence, a majority of the sentences might still be too easy for human listeners, in the sense that they could guess the right alternative based on the context. Therefore, we must select sentences that are not very common and cannot be easily predicted by humans. We do this by computing overlap between 4-grams in the language model training data and 4-grams in the sentence. We pick sentences with the lowest overlap as these sentences should be less predictable than sentences with high 4-gram overlap.

4. CASE STUDY: AUDIO-VISUAL SPEECH ENHANCEMENT EVALUATION

4.1. Audio-visual evaluation dataset

To create the AV dataset used in the evaluation, we selected a set of TED and TEDx videos[1] of public lectures delivered by a single speaker. Details about the train, dev and eval AV datasets can be found in [14]. After selecting the videos, we extracted sentences based on the manual transcriptions of the talks. We processed the data using a modified version of the lip-synchronisation pipeline [20] to extract a set of sentences in which the speaker face was visible and the shot was preserved throughout. The resulting dataset contains 1,389 extracted sentences from 30 speakers (15 females and 15 males).

To create the noisy mixes, we add an interferer (a competing speaker or noise) to the audio extracted from the videos using the frequency weighted SNR calculation method adopted in the Clarity Challenge [21]. The mixes were added at 3 SNR values per interferer. These values were chosen following the transcription task evaluation we describe in section 4.3.

4.2. General evaluation procedure

We present in the following subsections 3 different subjective evaluations. Before conducting these studies, we received ethical approval from the Ethics Committee in Edinburgh University. Participants in all studies were recruited through the Prolific Academic platform from a pool of native British speakers with self-reported normal hearing and normal (or corrected to normal) vision. Participants were asked to use a desktop machine or laptop, use headphones and to take the test in a quiet environment. Following the Hurricane Challenge [5], participants are presented with ordered mixes regarding interferer type and SNR with increasing level of difficulty throughout the test.

[1] https://www.ted.com/

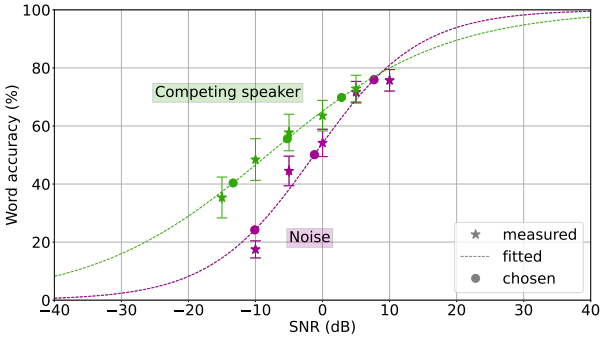

Fig. 2: Psychometric curves fitted on measured human word accuracy scores for competing speaker (green) and noise (purple). Error bars depict 95% confidence intervals.

Video clips can only be played once, therefore participants are instructed to pay close attention to each presented stimulus.

4.3. SNR calibration

We performed a preliminary listening test to find appropriate SNR values that would avoid ceiling or flooring effects. We first conducted an informal listening test with a native speaker. We presented video clips at different SNRs (distributed in 2 dB steps) that ranged from 0 db to -19 dB (competing speaker) and 0 dB to -15 dB (noise). We asked the participant to report those video clips in which it was possible to understand all or most of the words and those in which no words were intelligible. Based on these results, we selected SNR ranges from -10 dB to 10 dB for noise and -15 dB to 5 dB for competing speakers[2].

We then conducted a more formal evaluation to select 3 SNR values out of each range to use for the creation of the evaluation set. We did this following the procedure from the Hurricane Challenge [5] that estimates a psychometric curve based on the word accuracy scores obtained from a listening test. We collected responses from 40 participants. We presented participants a set of samples of a target speaker mixed with either noise or a competing speaker. Each video clip contained 7-10 words sentences. Participants were asked to type down the words they heard after watching each video. Figure 2 shows the curves obtained for each interferer. Based on these results we chose SNR values of -9.3, -1.2 and 6.9 dB (noise) and -13.5,-5.4 and 2.7 dB (competing speaker) that reflect word accuracy scores of 25%, 50% and 75% (noise) and 40%, 55%, and 70% (competing speaker).

4.4. Evaluation: transcription task versus proposed

To validate the method proposed in section 3, we conducted a study to compare intelligibility results (based on word accuracy scores) against a transcription-type test. From the original set of 1,389 sentences, we selected 120 sentences according to the method described in Section 3. The sentences were 7-10 words long. We presented the original (i.e., not enhanced) samples mixed with an interferer (half of the mixes have a competing

[2] These SNR ranges were used in the train and dev sets

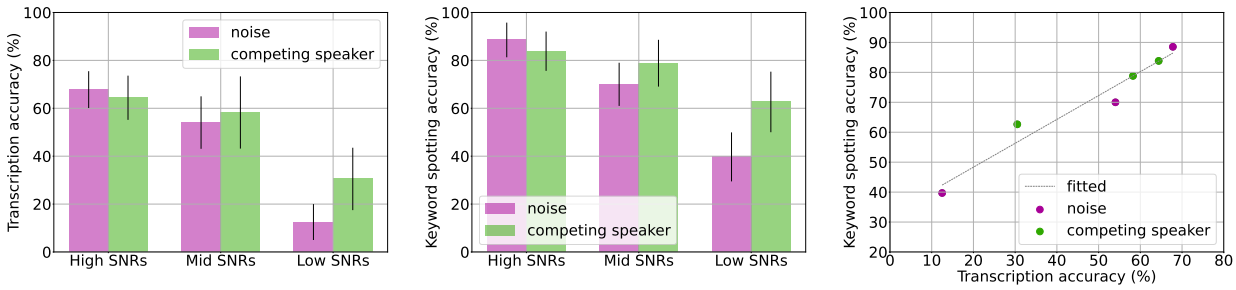

Fig. 3: Bar plots indicate word accuracy (%) per condition (SNR) and masker for the transcription-based task (left), and the keyword spotting method (middle). The plot on the right shows the correlation of word accuracy scores for both methods across conditions.

speaker scenario and the other half are mixed with noise). The noises belong to four categories that are a subset of the ones used in the train and dev sets, these are: microwave, washing-machine, hairdryer and soundscape.

We collected responses from 40 participants. 20 participants performed each evaluation (transcription versus keyword spotting). Both groups evaluated the same sentences in the same order. Figure 3 shows word accuracy scores obtained using each method. Results show a similar trend across the two methods: accuracy scores decrease in the more challenging conditions (low SNR), rankings according to the interferer type are the same and standard deviations calculated per SNR are similar in both evaluation methods. The keyword spotting method results in higher word accuracy by around (20 %), which reflects the chance level derived from the closed set of five possible responses. On average participants took 46 minutes to complete the study using the transcription method and 24 minutes using keyword spotting. A decrease in test duration is beneficial for two reasons: participants are less exposed to listening fatigue (that increases with longer listening tests) and less resources are required when collecting responses.

4.5. Evaluation of AVSE systems

After validating our proposed method we conducted a large-scale evaluation of speech enhancement systems submitted to [14]. We evaluated nine systems (including the baseline model), and the original (i.e., not enhanced) samples.

Participants were presented with the same 120 video clips described in the previous section. 95 participants took part in this study. Results are presented in Fig.4. Scores are significantly different from each other, except in system groups: (F, E, G), (G, H, I), (H, I), (J to B, C, D) and (B, C, D). While there isn't an obvious front-runner (i.e. the highest scoring system is not significantly different to all other systems), the paradigm was able to tell apart clear groups of systems, including those that are significantly worse than the non-enhanced data (A).

5. CONCLUSIONS

We presented a novel intelligibility evaluation method in which participants listen to a video clip and are then asked to select the word they heard out of a closed set of five alternatives. The list

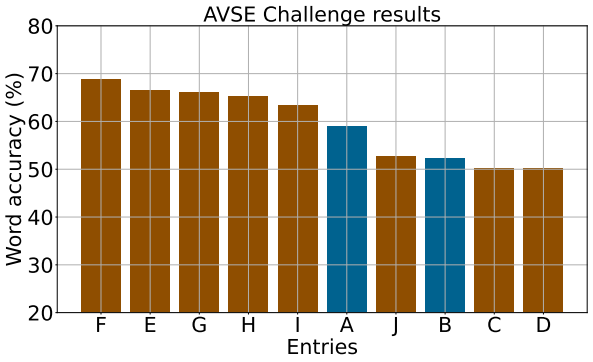

Fig. 4: Word accuracy (%) calculated across all maskers (LSD=3.35%). Original (A) and baseline (B) are in blue.

of alternatives comprises words that are phonetically similar to the target word plus a 'none of the above' option. Sentences, target words and alternatives were chosen according to phonetic distance and language model perplexity. We compared results from our proposed method against a transcription type task in terms of word accuracy scores and conclude that our method is an efficient and suitable alternative to intelligibility evaluation. We implemented our method in a large-scale evaluation of speech enhancement systems. We expect our work to layout a route on AV intelligibility evaluation methods that are suitable to assess performance of new AV speech technologies.

Acknowledgements This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) programme grant: COG-MHEAR (EP/T021063/1).

6. REFERENCES

[1] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, Dallas, USA, March 2010, pp. 4214–4217.

[2] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[3] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2153–2166, 2018.

[4] Y. Feng and F. Chen, "Nonintrusive objective measurement of speech intelligibility: A review of methodology," *Biomedical Signal Processing and Control*, vol. 71, pp. 103204, 2022.

[5] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Comm.*, vol. 55, no. 4, pp. 572–585, 2013.

[6] G. Llorach, F. Kirschner, G. Grimm, M. A. Zokoll, K. C. Wagener, and V. Hohmann, "Development and evaluation of video recordings for the olsa matrix sentence test," *International Journal of Audiology*, vol. 61, no. 4, pp. 311–321, 2022, PMID: 34109902.

[7] M. Gogate, K. Dashtipour, A. Adeel, and A. Hussain, "Cochleanet: A robust language-independent audio-visual model for real-time speech enhancement," *Information Fusion*, vol. 63, pp. 273–285, 2020.

[8] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 3244–3248.

[9] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, 2018.

[10] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Time domain audio visual speech separation," in *Proc. ASRU Workshop*, 2019, pp. 667–673.

[11] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *J. Sel. Topics in Sig. Proc.*, vol. 14, no. 3, pp. 530–541, 2020.

[12] J. Yu, S.-X. Zhang, B. Wu, S. Liu, S. Hu, M. Geng, X. Liu, H. Meng, and D. Yu, "Audio-visual multi-channel integration and recognition of overlapped speech," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 29, pp. 2067–2082, 2021.

[13] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-m. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, 03 2018.

[14] A. L. A. Blanco, C. Valentini-Botinhao, O. Klejch, M. Gogate, K. Dashtipour, A. Hussain, and P. Bell, "AVSE Challenge: Audio-visual Speech Enhancement Challenge," in *Proc. SLT Workshop*, 2022, pp. 465–471.

[15] S. Winters and D. Pisoni, "Speech synthesis, perception and comprehension of," in *Encyclopedia of Language & Linguistics (Second Edition)*, K. Brown, Ed., pp. 31–49. Elsevier, Oxford, second edition edition, 2006.

[16] W. A. Dreschler, *Hearing in the communication society D-2-2 deliverable*, 2006, http://hearcom.eu.

[17] C. Benoit, "An intelligibility test using semantically unpredictable sentences: towards the quantification of linguistic complexity," *Speech Comm.*, vol. 9, no. 4, pp. 293–304, 1990.

[18] IEEE, "IEEE recommended practice for speech quality measurement," *IEEE Trans. on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225 – 246, 1969.

[19] D. Kalikow, K. Stevens, and 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, pp. 1337, 1977.

[20] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 251–263.

[21] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. V. Muñoz, "Clarity-2021 Challenges: Machine Learning Challenges for Advancing Hearing Aid Processing," in *Proc. Interspeech*, 2021, pp. 686–690.