December 01, 2017

Quantifying speech intelligibility based on crowdsourcing

Edrei Chua

Jason Fedor

Caile Collins

Aaron Malenfant

Quantifying speech intelligibility based on crowdsourcing

ABSTRACT

The intelligibility of speech within media content, e.g., audio or video streams, is an important factor that determines the reach and popularity of the media. Objective measures of audio and speech quality, e.g., PESQ and SII scores, correlate poorly with human assessment. MOS, a widely accepted intelligibility test, is subjective, expensive, and time consuming.

Techniques disclosed herein provide an objective measure of the intelligibility of speech within video or audio content. Speech intelligibility scores are calculated based on the edit distance between human speech transcriptions of short clips and transcripts produced by an automatic speech recognizer. The speech intelligibility score is based on human rating and retains objectivity.

KEYWORDS

- Speech intelligibility
- Mean Opinion Score
- Edit distance
- PESQ
- SII
- Audio CAPTCHA

BACKGROUND

Speech intelligibility is an important property of videos and other media content, as it impacts the reach and popularity of such content. There are many factors that can affect the intelligibility of an audio clip, including signal-to-noise ratio, resolution of the audio, and the

accent, pronunciation, and tonal variation of a voice. It is a challenge to obtain a good measure of speech intelligibility.

Objective measures of audio and speech quality, such as the perceptual evaluation of speech quality (PESQ) and the speech intelligibility index (SII) do not directly measure speech intelligibility. Such measures correlate poorly with human assessment. Standardized listening tests based on human ratings are widely accepted for speech intelligibility. However, it is expensive to recruit subjects and takes substantial time to run listening tests in a lab environment. Besides, standardized listening tests, e.g., those using mean opinion score (MOS), are subjective because speech quality is ranked on a scale, e.g., a five-point scale, with the criteria for each quality category left to the listener.

DESCRIPTION

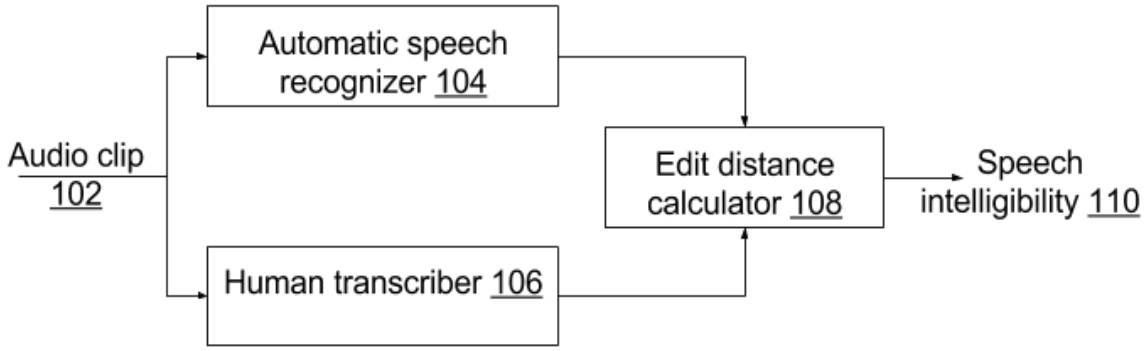

Fig. 1: Computing speech intelligibility

Fig. 1 illustrates the computation of speech intelligibility, per techniques of this disclosure. An audio clip (102) is transcribed (or captioned) using an automatic speech recognizer (104). The audio clip may originate from some larger media such as a longer audio or video stream.

Portions of the audio clip are served to human transcribers (106) who generate a transcript for that portion. These portions are short clips of audio selected uniformly and randomly from the original media. The short audio clips are restricted to only a few words such that the human transcribers do not infer words in the clips based on contextual knowledge, thus providing a more objective blind test.

An edit distance is calculated (108) between the two transcriptions. The edit distance between two transcriptions is defined, for example, as the minimum number of operations needed to transform one transcription to another. The speech intelligibility score (110) of the audio clip is calculated based on the edit distance. A low edit distance indicates high intelligibility, and vice-versa. A threshold is set for edit distance, and the short clips for which the edit distance threshold is not met are marked as intelligible.

A large number of short audio clips selected randomly from the original media are thus scored for speech intelligibility, using transcriptions from human transcribers. The overall speech intelligibility score of the original media is based on, for example, the arithmetic mean and standard deviation of the scores of the short clips taken from the original media. Overall speech intelligibility score can also be based on the percentage of clips marked as intelligible. Aside from edit distance, other measures of speech intelligibility include, e.g., the presence of particular words, homophones, etc., in both transcriptions.

In this manner, a measure of speech intelligibility is obtained that is both objective and obtained via human rating. The speech intelligibility score, as described herein, correlates well with human assessment. It is also relatively inexpensive to compute, especially when compared to standardized listening tests. For example, the human transcribers can be hired from web services that specialize in providing human contributors for various tasks, including speech

recognition. As a further example, the short audio clips can be presented as audio CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) where the challenge is set up as a speech transcription task for human transcribers.

The techniques described herein to compute speech intelligibility score have many applications. For example, a video-hosting web service could use the speech intelligibility score to rank videos or to filter out videos with poor speech intelligibility. When human transcribers are selected from a particular geographic region, the techniques can be used to test the intelligibility to people from particular parts of the world. The techniques can also be used to analyze trends on speech intelligibility variation across languages and demographics. The relationship between speech intelligibility and factors such as signal-to-noise ratio, resolution of the audio, and the accent, pronunciation, and tonal variation in a voice can be studied using the techniques.

When users permit, machine-learning models can be trained on crowd-sourced speech intelligibility scores, e.g., to automatically determine the degree of intelligibility of videos, and specifically, intelligibility to people from particular parts of the world. Audio and videoconferencing services can use the techniques described herein to assess the intelligibility of speech during an audio or video conference as a quality metric for the audio or video conference.

The performance of hearing aids or other devices can also be measured using the techniques of this disclosure. As such, the speech intelligibility score described herein can be used for applications where traditional speech intelligibility scores, e.g., PESQ, SII, MOS, etc., are used.

CONCLUSION

Techniques disclosed herein provide an objective measure of the intelligibility of speech within video or audio content. Speech intelligibility scores are calculated based on the edit distance between human speech transcriptions and transcripts produced by an automatic speech recognizer. The speech intelligibility score is based on human rating and retains objectivity. Human transcriptions are crowdsourced, and hence considerably cheaper than traditional standardized listening tests that are conducted in a laboratory.