July 31, 2017

Smart attenuation control for microphones

Sara Basson

Dimitri Kanevsky

Yiteng Huang

Follow this and additional works at: http://www.tdcommons.org/dpubs_series

Recommended Citation
Basson, Sara; Kanevsky, Dimitri; and Huang, Yiteng, "Smart attenuation control for microphones", Technical Disclosure Commons, (July 31, 2017)
http://www.tdcommons.org/dpubs_series/609

This work is licensed under a Creative Commons Attribution 4.0 License.
This Article is brought to you for free and open access by Technical Disclosure Commons. It has been accepted for inclusion in Defensive Publications Series by an authorized administrator of Technical Disclosure Commons.
Smart attenuation control for microphones

ABSTRACT

Microphone arrays are used to home into a source of sound. They find application in a variety of situations, e.g., conference rooms, video conferencing, virtual personal assistants, etc. When multiple speakers are participating in a joint conversation, microphone arrays tend to direct themselves to the loudest or clearest speakers. The problem with doing so is that the array does not provide timely and sufficient access to speakers with soft voices, accented speech, speech disorders, or whose speech is otherwise difficult to comprehend, e.g., low or unintelligible speech level. This disclosure describes techniques to identify people whose speech intelligibility levels are low, and to give them higher priority during selection of direction by the microphone array.

KEYWORDS

- Microphone array;
- Adaptive control;
- Speech intelligibility;
- Speech classification;
- Unintelligible speech

BACKGROUND

Microphone arrays follow one of several methodologies to home into audio from multiple speakers who participate in a joint conversation. Specifically, they are usually directed to the loudest source of speech. The problem with this approach is that the array may not provide timely and sufficient access to speakers whose voice is soft, or whose speech is difficult
to comprehend. Factors that make speech difficult to comprehend include: people with soft voices, people with accents, people with speech disorders, distance of the speaker from a microphone, people speaking near noise sources (e.g. people who sit near a window with noise from the street or near an air conditioner), etc. Directional microphones may have no time to switch between speakers in a quick interactive conversation, or may switch direction too late towards a person who has begun speaking, thereby failing to capture the beginning of a spoken utterance for such a speaker. This problem is more acute for omni-directional or non-array microphones, which cannot specify listen-direction.

DESCRIPTION

This disclosure describes techniques to identify people whose speech is difficult to comprehend, and to give them higher priority when listen-directions are set for microphones used with user consent and permission. The identification of people with speech of low intelligibility is done via speech classification by speech intelligibility level. The listen-directions for microphones are chosen via methods of automatic adaptive control.

A speech intelligibility score is created based on several variables, e.g., the characteristics of the speaker; the material used for speaking and its mode of transmission (e.g. audio or video recordings versus live presentations); the characteristics of the listeners (e.g. are they familiar or unfamiliar with the speakers?); etc. Characteristics of the speaker that influence intelligibility include segmental (e.g., individual sounds) factors, suprasegmental factors (e.g. rate, intonation and rhythm of speech), speech sound production, etc. Segmental characteristics strongly associated with reduced speech intelligibility include omission of word initial phonemes, voicing errors, errors of consonant clusters, consonant substitutions, soft voices, accents and unidentifiable distortions.
Fig. 1: Block diagram of an example machine-learned model

Fig. 1 depicts a block diagram of an example machine-learned model according to example implementations of the present disclosure. As illustrated in Figure 1, in some implementations, the machine-learned model is trained to receive input data of one or more types and, in response, provide output data of one or more types. Thus, Fig. 1 illustrates the machine-learned model performing inference.

In some implementations, the input data can include one or more features that are associated with an instance or an example. In some implementations, the one or more features associated with the instance or example can be organized into a feature vector. In some implementations, the output data can include one or more predictions. Predictions can also be referred to as inferences. Thus, given features associated with a particular instance, the machine-learned model can output a prediction for such instance based on the features. As examples, the input data includes speech captured by the microphone array with the consent and permission of the speakers.

In some implementations in which the machine-learned model performs classification, the machine-learned model can be trained using supervised learning techniques. For example, the machine-learned model can be trained on a training dataset that includes training examples labeled as belonging (or not belonging) to one or more classes.
Fig. 2 illustrates how a suitable training dataset is created. A transcription service receives speech, anonymized if necessary, of particular speakers with the consent and permission of the speakers. The transcription service could be performed by human transcribers or by automatic speech recognition systems. The output of the transcription service is transcribed text. If the transcribed text contains markers of unrecognizable speech, e.g., “inaudible”, “noise”, “multiple speakers”, etc., then that is an indication of speech that is less than fully intelligible. A large number of “inaudible”, “noise”, “multiple speakers”, etc. labels would indicate low speech intelligibility.

If the transcription service is provided by an automatic speech recognition system then one can use confidence metrics that approximately define a quality of a decoded transcript. A higher than an average number of decoding errors for a speaker may indicate low speech intelligibility for that speaker. In this manner, speech samples and speakers are identified who have low speech intelligibility. Such a set of speech samples serve as a training dataset for a machine-learned model. Another approach to obtain a training set for a machine-learned model is to create a corpus of speech with various characteristics (soft speech, accented speech, etc.)
and combine these speech data with various types of noises. Human evaluators could listen to these types of speech, anonymized if necessary, with the consent and permission of the speakers, and rank the types of speech by level of intelligibility. This dataset of speech and classification labels is used to train the machine-learned model.

![Diagram: Labeled training set](image)

**Fig. 3: Training a machine-learned model with speech samples of low intelligibility**

This is illustrated by the example in Fig. 3 where a labeled training set comprising speech samples that are soft (306), accented (308), noisy (310), etc. is fed to a machine-learned model in order to train it.

In operation, speakers with low speech-intelligibility are identified using a trained, machine-learned model, as obtained using training techniques described above. In some implementations, the machine-learned model can be or include one or more artificial neural networks (also referred to simply as neural networks). A neural network can include a group of connected nodes, which also can be referred to as neurons or perceptrons. A neural network can be organized into one or more layers. Neural networks that include multiple layers can be referred to as “deep” networks. A deep network can include an input layer, an output layer, and
one or more hidden layers positioned between the input layer and the output layer. The nodes of the neural network can be connected or non-fully connected.

In some implementations, the machine-learned model can be or include one or more feed forward neural networks. In feed forward networks, the connections between nodes do not form a cycle. For example, each connection can connect a node from an earlier layer to a node from a later layer.

In some instances, the machine-learned model can be or include one or more recurrent neural networks. In some instances, at least some of the nodes of a recurrent neural network can form a cycle. Recurrent neural networks can be especially useful for processing input data that is sequential in nature. In particular, in some instances, a recurrent neural network can pass or retain information from a previous portion of the input data sequence to a subsequent portion of the input data sequence through the use of recurrent or directed cyclical node connections. Sequential input data can include words in a sentence, e.g., for natural language processing, speech detection or processing, etc. Example recurrent neural networks include long short-term (LSTM) recurrent neural networks; gated recurrent units; bidirectional recurrent neural networks; continuous time recurrent neural networks; neural history compressors; echo state networks; Elman networks; Jordan networks; recursive neural networks; Hopfield networks; fully recurrent networks; sequence-to-sequence configurations; etc.

In some implementations, the machine-learned model can be or include one or more convolutional neural networks. In some instances, a convolutional neural network can include one or more convolutional layers that perform convolutions over input data using learned filters. Filters can also be referred to as kernels. Convolutional neural networks can also be applied for natural language processing. In some implementations, the machine-learned model
can be or include one or more other forms of artificial neural networks such as, for example, deep Boltzmann machines; deep belief networks; stacked autoencoders; etc. Any of the neural networks described herein can be combined (e.g., stacked) to form more complex networks.

In some implementations, one or more neural networks can be used to provide an embedding based on the input data. For example, the embedding can be a representation of knowledge abstracted from the input data into one or more learned dimensions. In some instances, embeddings can be a useful source for identifying related entities. In some instances embeddings can be extracted from the output of the network, while in other instances embeddings can be extracted from any hidden node or layer of the network (e.g., a close to final but not final layer of the network).

Once trained, the neural or other machine-learned model directs the microphone array by classifying directions by their levels of speech intelligibility. The array’s directional sensitivity is prioritized towards speech sources with less intelligibility. In case of microphones that are only uni- or bi-directional, directional sensitivity can be achieved by quickly rotating such microphones towards a high-priority (unintelligible speech) direction.

In situations in which certain implementations discussed herein may collect or use personal information about speakers or users (e.g., user data, user’s speech, information about a user’s social network, user's location and time at the location, user's biometric information, user's activities, user’s online history and demographic information), users are provided with one or more opportunities to control whether information is collected, whether the personal information is stored, whether the personal information is used, and how the information is collected about the user, stored and used. That is, the systems and methods discussed herein collect, store and/or use user personal information specifically upon receiving explicit
authorization from the relevant users to do so. For example, a user is provided with control over whether programs or features collect user information about that particular user or other users relevant to the program or feature. Each user for which personal information is to be collected is presented with one or more options to allow control over the information collection relevant to that user, to provide permission or authorization as to whether the information is collected and as to which portions of the information are to be collected. For example, users can be provided with one or more such control options over a communication network. In addition, certain data may be treated in one or more ways before it is stored or used so that personally identifiable information is removed. As one example, a user’s identity may be treated so that no personally identifiable information can be determined. As another example, a user’s geographic location may be generalized to a larger region so that the user's particular location cannot be determined.

CONCLUSION

A microphone array used, for example, in a video conferencing or virtual assistant setting, often directs itself towards speakers with the loudest or clearest voices. By doing so, people with soft voices, accented speech, speech disorders, or otherwise having low speech intelligibility are often ignored by the microphone array. Techniques disclosed herein address this problem by setting directional focus for microphones based on intelligibility, rather than loudness, of speech. Transcriptions provided by human transcribers, or confidence metrics/transcription error rates provided by automatic speech recognizers, are used to measure speech intelligibility. Such transcriptions and confidence metrics are used with corresponding speech samples to train a machine learner. Once trained, the machine learner directs a microphone array (or rotates a uni-directional microphone) based on speech intelligibility.