Technical Disclosure Commons

Defensive Publications Series

July 24, 2017

Generation of speech training data for special speech recognition tasks

Dimitri Kanevsky

Andrew Senior

Sara Basson

Generation of speech training data for special speech recognition tasks

ABSTRACT

Speech recognition is widely used as a voice-user interface in several settings, e.g., interactive voice response, virtual personal assistants, transcription, translation applications, etc. Although speech recognition technology has advanced far enough to be useful to a sizeable number of human speakers, there are still populations that cannot take full advantage of speech recognition. For example, people with impaired speech, speakers of rare languages or dialects, with strong accents, etc. have difficulty using an application that uses speech recognition. The reason for such user difficulty is that there is insufficient data to train an automatic speech recognizer to recognize such relatively rare speech. This disclosure describes techniques for creating a large training set out of a small set of speech samples. Acoustic and linguistic features peculiar to a class of speakers are extracted out of a small set of their speech samples, with their consent and permission. These features presented as constraints to a speech synthesizer in order to generate a larger training set.

KEYWORDS

- Non-standard speech
- Impaired speech
- Accented speech
- Speech recognition
- Training-set generation
- Speech synthesis

BACKGROUND

There are populations of human speakers whose speech is difficult to be processed by automatic speech recognizers. These are, for example, populations that speak rare languages, unusual dialects, have impaired speech (e.g., those who've suffered strokes, who stutter, or who have deaf characteristics in their speech), have strong accents etc. At the same time, voice interface to the computer is becoming increasingly popular, e.g., in the form of virtual personal assistants; yet it is not accessible to the above-described classes of users. The class of people with impaired speech is estimated as being tens of millions in the United States alone. If the voice-user interface is not tailored to this class of people, then they will be shut out from using this emerging and potentially pervasive technology.

A principal challenge in getting automatic speech recognizers to recognize speech of a specific, "non-standard" type, e.g., from populations as listed above, is to get a training dataset large enough to train the speech recognizer. An automatic speech recognizer, based for example on neural networks, requires thousands of hours of voice data for training purposes. It is difficult to collect such large amounts of data for each distinct type of speech disorder, thick accent or rare language.

DESCRIPTION

This disclosure describes techniques to efficiently generate training data so as to enable training of speech recognizers in non-standard speech types. A small, representative sample of the non-standard speech type, e.g., comprising a particular speech impairment such as Broca's aphasia, is used to create a set of characterizing acoustic features. These features are then placed as constraints to a speech synthesizer, which generates a larger set of raw-audio speech samples with characteristics closely corresponding to the original, smaller set of speech samples. This larger set of speech samples serves as a data corpus to train speech recognizers to recognize speech originating from a speaker of the non-standard speech type.

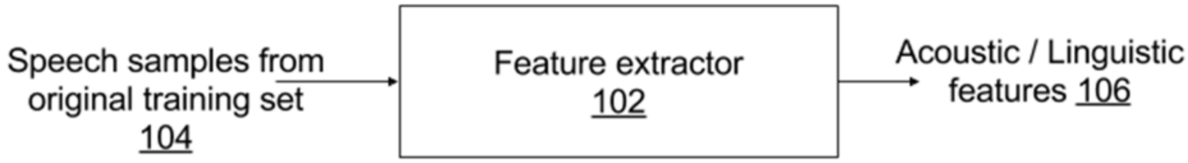

Fig. 1: Analyzing speech samples of a smaller ("original") training set in order to obtain acoustic and/or linguistic features

Fig. 1 illustrates an example process by which acoustic and/or linguistic features are extracted out of samples of a non-standard speech type. A feature extractor (102) processes speech samples from an original training set (104) and extracts acoustic and/or linguistic features (106). The speech samples 104 are "original" in the sense that they are samples from actual non-standard human speakers of a particular type (e.g., accented, rare language, speech impaired, deaf characteristics, etc.), where the samples are taken with the speakers' consent and permission, and anonymized.

As explained previously, such samples taken directly from human speakers are typically of small size, generally not enough to train a speech recognizer. The output of the feature extractor is a set of acoustic features, represented mathematically by a set of vectors, e.g., x_1, x_2 $, \dots , x_t$. Linguistic features, denoted L_1, $L_2 , \dots , L_t$ are associated with the acoustic features. The association of acoustic with linguistic features is represented by writing associated features in pairs, e.g., $(x_1 , L_1) ; (x_2 , L_2) ; \dots ; (x_t , L_t)$. As an example, if a standard speaker of English utters the sentence "I go home," then L_1 equals "I", L_2 equals "go", and L_3 equals "home." A person with certain types of aphasia might say "I … home," where the ellipsis denotes a gap in speech.

In this case, the linguistic features are as follows: L_1 equals “I”, L_2 equals “...”, and L_3 equals “home.”

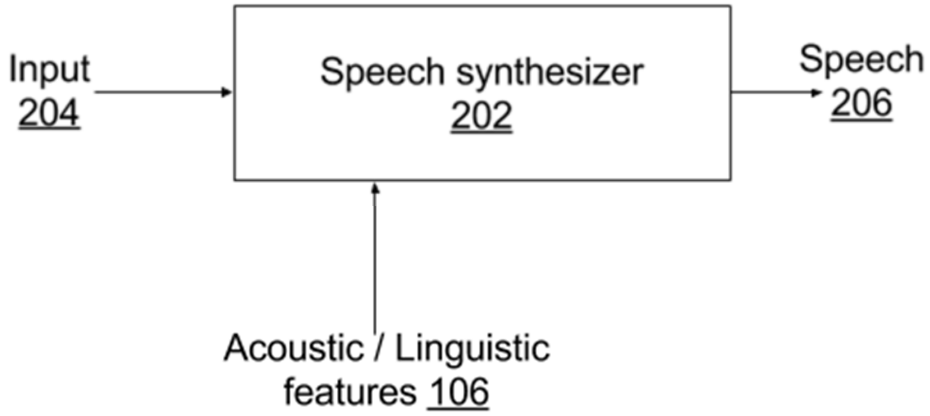

Fig. 2: Generating a larger training set using a synthesizer

Fig. 2 illustrates the generation of a larger training set using a speech synthesizer (202). The speech synthesizer is fed with an input (204), and it synthesizes speech (206) subject to constraints established by the just-found acoustic/linguistic features (106) of the non-standard speech. The input 204 could be, for example, text, or it could be speech samples from the original training set.

An illustrative example of a speech synthesizer is the Wavenet speech synthesizer[1]. Wavenet uses deep neural networks to generate raw audio waveforms. Wavenet allows creation of speech samples satisfying some constraints, e.g., linguistic or acoustic, derived from a small number of speech samples.

This set of speech samples generated by speech synthesizer 202, e.g., a Wavenet speech synthesizer, is tailored to the specifics of the non-standard speech type, since it is generated under constraints originating from features of the non-standard speech. Speech samples generated by speech synthesizer 202 can be used as a speech data corpus to train acoustic models for speech recognition purposes. For example, “stutter” features (repetitions, short/long pauses) can be added as a linguistic constraint into a general set of linguistic features that are used to condition the generation of audio features for training acoustic models.

Other speech disorders could also be represented as specific linguistic constraints. For example, after strokes, people with Broca's aphasia may have language that is reduced to disjointed words. Sentence construction is poor, and the speaker omits function words and inflections (bound morphemes). A person with expressive aphasia might say "Daughter ... College ... Smart ... Good ... Good." Content words, e.g., nouns, verbs, etc., may be used in speech, but sentences are difficult to produce due to problems with grammar, resulting in "telegraphic speech." Again, "expressive aphasia" could be added as a linguistic constraint to the speech synthesizer in order to generate speech, and hence training data, that is representative of speech affected by that condition.

New virtual speakers could be generated by embedding multiple speakers into the synthesizer. Varying the embedding level of each constituent speaker would generate a variety of new speakers. A virtual speaker's label, e.g., constraint, may be augmented with a certain condition, e.g., stutter, Broca's aphasia, etc., so that training may be performed on multi-speaker data. By turning on or off a virtual speaker's constraint one could generate speech with and without impairment, accent, etc. Thus, one synthesizes non-standard speech as it would be spoken by individuals with a speech condition as if they didn't have that condition.

Similar to machine translations between natural languages, one could use the framework described herein to automatically learn translations within the linguistic domain. For example, using a seq2seq model one could transform stuttered speech to fluent speech prior to synthesis. The opposite process is operative during speech recognition; for example, stuttered speech is decoded to fluent speech using, for example a maximum a posteriori method.

In another scenario, a large training set could be created by randomly transforming existing training data, with constraints on the transformations that reflect specific speech conditions.

Techniques described herein, e.g., the generation of a large set of speech samples with a specific non-standard speech condition, the mixing of speech with and without impairment, the translation of speech between impaired and fluent, etc., enable speech recognition developers to generate training data with specific linguistic or acoustic characteristics. The techniques also allow control of the amount of each such characteristic in the training data.

Where neural networks are used in this disclosure, e.g., in the speech synthesizer, etc., the neural network includes a group of connected nodes, 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. Other types of neural networks, e.g., feed-forward neural networks, recurrent neural networks, convolutional neural networks, deep Boltzmann machines, deep belief networks, stacked autoencoders, etc. can be used. Any of the neural networks listed above can be combined (e.g., stacked) to form more complex networks.

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

Automatic speech recognition systems have difficulty recognizing “non-standard” speech, e.g., speech with impairments, thick accents, etc. Speech recognizers do not work well with non-standard speech types due to the lack of sufficiently large amounts of non-standard training speech data. Techniques of this disclosure enable the creation of a large training dataset having a specific speech condition (e.g., speech impairment, accent, etc.) from a small initial training set. The small initial training set is analyzed in order to extract features (or models) characteristic to the speech condition. These features are used to constrain a speech synthesizer as it generates speech. The output of the speech synthesizer serves as a large training dataset for an automatic speech recognizer. Additionally, techniques disclosed herein enable the generation of speech with and without a specified speech condition. Furthermore, translations can be performed within the linguistic domain, e.g., transform from stuttered to fluent speech, etc.

REFERENCES

[1] “WaveNet: A Generative Model for Raw Audio,” A. v. d. Oord *et al.*, arXiv:1609.03499 [cs.SD]