Technical Disclosure Commons

Defensive Publications Series

May 2020

Utterance Augmentation for Speaker Recognition

Jin Shi

Quan Wang

Yeming Fang

Gang Feng

Zhengying Chen

See next page for additional authors

Inventor(s)
Jin Shi, Quan Wang, Yeming Fang, Gang Feng, Zhengying Chen, Jason Pelecanos, Ignacio Lopez Moreno, Andrea Chu, and Pedro Moreno Mengibar

Utterance Augmentation for Speaker Recognition

ABSTRACT

The speaker recognition problem is to automatically recognize a person from their voice. The training of a speaker recognition model typically requires a very large training corpus, e.g., multiple voice samples from a very large number of individuals. In the diverse domains of application of speaker recognition, it is often impractical to obtain a training corpus of the requisite size. This disclosure describes techniques that augment utterances, e.g., by cutting, splitting, shuffling, etc., such that the need for collections of raw voice samples from individuals is substantially reduced. In effect, the original model works better on the augmented utterances on the target domain.

KEYWORDS

- Speaker recognition
- Utterance augmentation
- Synthetic training data
- Training corpus
- Smart speaker
- Smart display
- Virtual assistant
- Voice assistant

BACKGROUND

The speaker recognition problem is to automatically recognize a person from their voice. Speaker recognition has applications in several domains, e.g., for user authentication; in virtual assistant software and devices such as smart speakers, smart displays, and other voice-activated appliances; in call centers, etc. The training of a speaker recognition model typically requires a large training corpus, e.g., multiple voice samples from a very large number of individuals. In the diverse domains of application of speaker recognition, it is often impractical to collect a training corpus of such size.

Prevailing speaker recognition models map utterances into vectors and provide a distance measure between vectors. Vectors of utterances from the same speaker are close to each other while vectors of utterances from different speakers are far away from each other. Some techniques modify the vector of utterances directly; however, such modifications to vectors can be difficult to interpret. Other techniques attempt to synthesize training data using adversarial training of speaker recognition models.

DESCRIPTION

This disclosure describes techniques that augment utterances, e.g., by cutting, splitting, shuffling, etc., such that the need for collections of raw voice samples from individuals is substantially reduced. With such augmentation, the original model works better on the target domain.

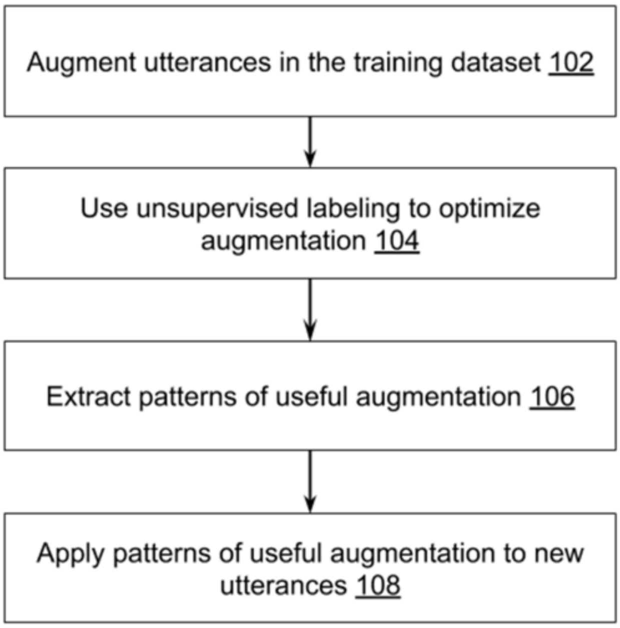
Augment utterances in the training dataset 102
Use unsupervised labeling to optimize augmentation 104
Extract patterns of useful augmentation 106
Apply patterns of useful augmentation to new utterances 108

Fig. 1: Utterance augmentation for speaker recognition

Fig. 1 illustrates the components of utterance augmentation for speaker recognition, per the techniques of this disclosure.

Utterance augmentation (102): Utterances in a relatively small training dataset are augmented. An utterance can be augmented by various techniques, e.g., cutting, splicing, shuffling, etc. Within each technique of augmentation, the parameters of augmentation can vary widely. For example, in a cutting-style augmentation, an utterance can be cut at any point along its length. As another example, in a splicing-style augmentation, an utterance can be divided into many sections and spliced back together in a number of different ways.

Unsupervised labeling (104): Unsupervised labeling is used to optimize an augmentation, e.g., to determine the cut-point for a cutting-style augmentation. As explained before, utterances are mapped to vectors, with a distance notion $dist(x, y)$ defined between vectors x and y. Let $X(s)$

denote the set of vectors of utterances from a speaker *s* that are in the dataset. The loss function *L(s)* minimized by the speaker recognition model can be written as:

$$L(s) = -\alpha \sum_{x,y \in X(s)} dist(x,y) + \beta \sum_{x \in X(s), y \notin X(s)} dist(x,y)$$

where **𝞪** and ***β*** are constants. An augmentation to an utterance *x∈X(s)* creates a new vector *x'*, with the change in loss given by:

$$\Delta L(s,x,x') = -\alpha \sum_{y \in X(s), y \neq x} (dist(x',y) - dist(x,y)) + \beta \sum_{y \notin X(s)} (dist(x',y) - dist(x,y))$$

With the above change-in-loss function, unsupervised labeling effectively evaluates the benefit or cost accrued by a given augmentation. For example, in cutting-style augmentation, which discards the section of the utterance after the cut point, utterances that are cut too short cause a loss in performance. On the other hand, cutting fifty milliseconds off a two-second utterance does not result in a perceivable loss in performance. In any case, the above formula predicts the benefit or cost for each cut-point.

Pattern extraction (106): Although unsupervised learning results in augmentation instances with labels that indicate the benefit or cost from a given augmentation to speaker-recognition quality, the labels result from a relatively small dataset. To generalize the labeling results from unsupervised learning, patterns of useful augmentation are extracted from the totality of labeled augmentation instances. The patterns can be extracted using, for example, rules or machine learning models. For example, in cutting-style augmentation, for each cut-point, a spectrogram is generated for a time window before and after the cut-point, so that a two-dimensional array context is attached to each cutting instance. A classifier, for example, a convolutional neural network, can be trained on cutting contexts and labels.

Pattern application (108): Patterns extracted during pattern extraction are applied to new utterances. For example, in cutting-style augmentation, new utterances can be generated using the cutting context discovered during pattern-extraction. The benefit or cost for a given cut-point can be inferred, a decision made whether to cut or to not, and, based on classifier predictions, the position of the cut-point (if any).

Per the techniques, the speaker recognition model is left unchanged, as are the vectors that the utterances map to. The synthesized data generated using the described techniques avoids excessive noise and makes full use of the raw training data, e.g., training data obtained from real speaker's utterances. The augmentation techniques result in quicker response and better accuracy of a speaker recognition model. The techniques apply to both phases of speaker recognition, e.g., the enrollment phase, a relatively infrequent phase where vectors for a given speaker are obtained, and the identification phase, a relatively frequent phase where a new utterance is compared with enrolled speakers to determine speaker identity.

CONCLUSION

The training of a speaker recognition model typically requires a very large training corpus, e.g., multiple voice samples from a very large number of individuals. In the diverse domains of application of speaker recognition, it is often impractical to obtain a training corpus of the requisite size. This disclosure describes techniques that augment utterances, e.g., by cutting, splitting, shuffling, etc., such that the need for collections of raw voice samples from individuals is substantially reduced. In effect, the original model works better on the augmented utterances on the target domain.

REFERENCES

[1] Wang, Qing, Wei Rao, Sining Sun, Leib Xie, Eng Siong Chng, and Haizhou Li. “Unsupervised domain adaptation via domain adversarial training for speaker recognition.” In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4889-4893. IEEE, 2018.

[2] Chien, Jen-Tzung, and Kang-Ting Peng. “Adversarial Learning and Augmentation for Speaker Recognition.” In *Proc. Odyssey: The Speaker & Language Recognition Workshop*, pp. 342-348. 2018.