SUFFICIENCY QUANTIFICATION FOR SEAMLESS TEXT-INDEPENDENT SPEAKER ENROLLMENT

Gokcen Cilingir, Jonathan Huang, Mandar S Joshi, Narayan Biswal

Intel Corporation

ABSTRACT

Text-independent speaker recognition (TI-SR) requires a lengthy enrollment process that involves asking dedicated time from the user to create a reliable model of their voice. Seamless enrollment is a highly attractive feature which refers to the enrollment process that happens in the background and asks for no dedicated time from the user. One of the key problems in a fully automated seamless enrollment process is to determine the sufficiency of a given utterance collection for the purpose of TI-SR. No known metric exists in the literature to quantify sufficiency. This paper introduces a novel metric called phoneme-richness score. Quality of a sufficiency metric can be assessed via its correlation with the TI-SR performance. Our assessment shows that phoneme-richness score achieves -0.96 correlation with TI-SR performance (measured in equal error rate), which is highly significant, whereas a naive sufficiency metric like speech duration achieves only -0.68 correlation.

Index Terms— Speaker recognition, text-independent, seamless enrollment, sufficiency metric, phoneme-richness metric

1. INTRODUCTION

As speech recognition technology is getting better with the advances in deep learning techniques, speech is getting closer to becoming a completely pervasive user interface in the human-computer interaction landscape. Delivering secure system interaction becomes an important part of the puzzle and speaker recognition is a natural fit to deliver secure system interaction. Speaker recognition techniques make use of machine learning to determine the identity of an enrolled speaker from a segment of speech uttered by them.

Two types of speaker recognition (SR) can be defined in terms of the constraints on the content of speech required for recognition. For text-dependent (TD) SR, the words used to enroll and test should be the same. TD-SR can be used with wake up phrases and with longer, more phonetically rich pass-phrases. The enrollment process requires a few repetitions of the same phrase. Text-independent (TI) SR, on the other hand, aims to recognize speaker identity with no constraints on the speech content, which makes it possible to recognize speakers during natural conversational speech [1]. TI-SR requires a lengthy enrollment process that involves asking dedicated time from the user to create a reliable model of their voice. It is important to capture a persons voice uttering a wide range of different speech content to ensure the robustness of the resulting speaker model and therefore special phoneme-rich passages are often required to be read. In general, TI-SR requires longer enrollment and test utterances to achieve the error rates TD-SR achieves.

Usability of TI-SR can be greatly improved by a system design that offers a seamless enrollment process where data collection for enrollment happens in the background. [2] discusses a seamless enrollment system where face recognition and lip reading detection are used to determine when a target speaker is talking, while capturing speech in the background for the purpose of TI-SR training.

When the enrollment process is not controlled by asking a phoneme-rich passage to be read, one requires a metric to analyze the value of an utterance pool with respect to its ability to inform the speaker model training process to create robust speaker models. No known metric exists in the literature for quantifying the value/sufficiency of an utterance pool for the purpose of TI-SR enrollment. In addition to making seamless enrollment more robust, sufficiency metrics provide useful information to improve traditional TI-SR enrollment and identification/verification processes. An SR system may utilize sufficiency metrics to determine optimal enrollment durations in order to achieve target accuracies. Sufficiency metrics can be utilized in confidence quantification for multi-session enrollment [3] and model fusion [4, 5]. Sufficiency metrics can also be used during verification/identification for confidence modeling. A higher sufficiency score over a test utterance may indicate a lower expected error rate and therefore higher confidence in the system decision.

This paper introduces a novel sufficiency metric called phoneme-richness score. Quality of a sufficiency metric can be assessed via its correlation with the TI-SR performance. Our assessment shows that phoneme-richness score achieves -0.96 correlation with TI-SR performance, which is highly significant, whereas a naive sufficiency criteria like speech duration achieves only -0.68 correlation.

Literature exists on analyzing the quality of a speech signal in terms of distortions due to channel effects and envi-

ronmental noise. Quality measures include noise classification, signal to noise ratio, and universal background model (UBM) score [6]. [7, 8, 9, 10, 6, 11] describe ways to define such quality measures and utilize them to improve the robustness and performance of speaker recognition. These quality measures mainly focus on quantifying the cleanliness of a speech signal from distortions, rather than sufficiency of a set of speech signals for the purpose of robust text-independent enrollment. Although phoneme-richness score can also be utilized as a speech quality metric, this paper mainly focuses on its use as a sufficiency metric for robust text-independent enrollment. In Section 2, we briefly describe a high-level system design for a seamless enrollment process to show where sufficiency quantification fits in the big picture. In Section 3, we explain how we define phoneme-richness score. In Section 4, we present the details of our experiments to showcase the utility of phoneme-richness score as a sufficiency metric for robust text-independent enrollment. We conclude with a summary in Section 5.

2. SEAMLESS ENROLLMENT FOR TEXT-INDEPENDENT SPEAKER RECOGNITION

In this section, we will briefly cover the high-level flow of the seamless enrollment process, given in Figure 1. This design for seamless enrollment for TI-SR requires that there exists at least one other tool in the system to predict speaker identity during data collection process for TI-SR enrollment. One can use the wake up phrase and apply text-dependent SR to predict speaker identity. Other biometrics on face, iris, body, wearable sensors, etc. can be used for identity detection. [2] uses face recognition and lip reading detection to determine speaker identity in their design.

In this design, data collection is assumed to happen continuously with an always on device (like Amazon Echo or Google Home) and the data is properly tagged and stored in an utterance pool. The utterance pool is processed at certain time intervals with the aim of either adopting/updating an already created text-independent speaker model or creating a new speaker model. All the utterances collected from a target speaker is subjected to a process called sufficiency quantification. The aim of this process is to determine the value of an utterance pool for the purpose of TI-SR enrollment. A sufficiency metric can be based on quantifying diversity, quantity, and predictive quality. Multiple sufficiency metrics may be combined to make a stronger metric. Diversity in the context of speech is defined in terms of phoneme diversity and will be referred as phoneme-richness. A direct way to quantify phoneme-richness is to use phoneme predictors that are utilized in speech recognition. In the rest of the paper, we will discuss how we defined phoneme-richness independent of phoneme predictors, utilizing a Gaussian mixture model (GMM).

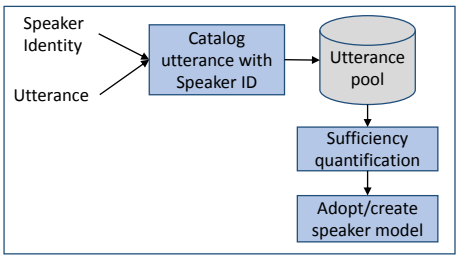

Fig. 1. Seamless enrollment/adaptation flow

3. PHONEME-RICHNESS QUANTIFICATION

We extract 26 log mel-filterbank energies as features from overlapping windows with length 25 msec and overlap amount of 10 msec. The FFT size was 512 and we applied preemphasis filter with 0.97 preemph. We train a 256-mixture GMM that we call the Universal Background Model (UBM), using the train portion of the TIMIT dataset [12], which contains clean voice recordings of 462 speakers, each uttering 10 phoneme-rich sentences (each containing 7-10 words).

Phoneme distribution for a given utterance pool is represented as a histogram over the highest-scoring Gaussian mixture indices in the UBM. A reference histogram is calculated using the test portion of the TIMIT dataset. This histogram represents the ground truth for an adequately phoneme-rich utterance pool. Test portion of the TIMIT dataset contains clean voice recordings of 226 speakers, each uttering 10 phoneme-rich sentences, which can be considered an adequately phoneme-rich utterance pool for our purposes. Figure 2 illustrates the offline process where the TIMIT dataset is used to create the reference histograms. An energy-based Voice Activity Detection (VAD) technique is used to separate speech content from the non-speech. Energy is calculated over an overlapping window with length 25 msec and overlap amount of 10 msec. Noise floor is calculated per recording by taking the average energy in the first 10 frames. Noise floor adjusted energy for each frame is thresholded to identify high energy frames, which are assumed to contain speech content. Since TIMIT contains clean recordings with no additional noise, energy-based VAD is sufficient.

In addition to creating a reference histogram for the speech content, we create a histogram over the non-speech content. We only use this histogram to determine the bins most populated by non-speech content. We then eliminate these bins in the speech reference histogram so they do not affect phoneme-richness score calculations. The objective is to minimize the effect of non-speech content in the creation of reference speech histogram.

Figure 3 shows the process applied when an utterance pool in question requires phoneme-richness quantification. Speech and non-speech histograms are extracted from the utterance pool in the same way reference histograms are created. Normalization is required before calculating distance

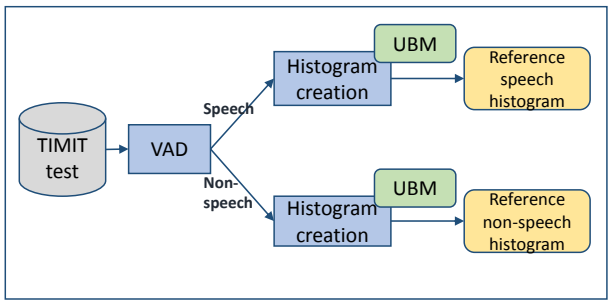

Fig. 2. Reference histogram creation process

between histograms and can be achieved by dividing each value in the histogram by the sum of the values.

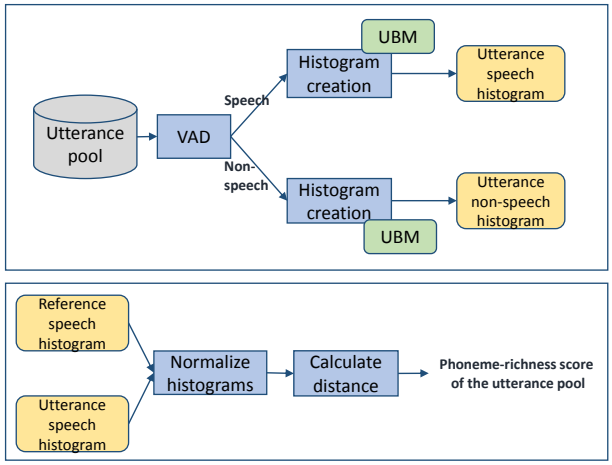

Fig. 3. Phoneme-richness score calculation process for a given utterance pool

For distance calculation between a reference histogram and a histogram representation of an utterance pool, a custom distance metric is applied. The bins in the utterance histogram that have a smaller value than the corresponding bins in the reference histogram are identified. The distance score is calculated as the sum of the bin value differences for the identified bins. The objective is to eliminate the effect of an overrepresented phoneme set in an utterance pool over the distance score. Distance score only sums up the effect of underrepresented phoneme sets in an utterance pool. The distance score is subtracted from 1 to achieve the phoneme-richness score. This works since range of distance score is 0 to 1. Phoneme-richness score P is calculated using the equations given below, where reference histogram is denoted by H_{ref} and utterance pool histogram is denoted by H_{pool}:

$$H_{ref} = hr_0, hr_1, ..., hr_n \quad (1)$$

$$H_{pool} = hp_0, hp_1, ..., hp_n \quad (2)$$

$$P = (1 - \sum_{i}^{n} hr_i - hp_i \mid hr_i > hp_i) \quad (3)$$

4. RESULTS

We used a proprietary dataset containing utterances from 40 speakers (gender-balanced), sampled at 16 kHz. Each speaker uttered 180 short commands with varying lengths, 10 repeats of 20 trigger words and a phoneme-rich passage. We separated 180 short command utterances randomly into 25 batches to experiment with adding phonetic richness by incrementing speech data one batch at a time. Each batch contains about 3 seconds of speech duration determined by using an energy-based VAD. One batch contains on average 5.5 commands. Each command contains on average 0.5 seconds of speech content. We devised several enrollment scenarios to showcase the metric behavior:

- Enrollment with 10 repeats of the trigger word hello computer (average speech duration: 2.7 seconds)
- Enrollment with 10 repeats of 4 trigger words (average speech duration: 12 seconds)
- Enrollment with the phoneme rich passage (average speech duration: 11.8 seconds)
- 1 to 15 utterance batches are used for enrollment. (average speech duration for each batch: 5.5 seconds)

We used on average 55 short utterances for target trials per speaker and on average 11 utterances for each imposter per speaker. For each speaker, all the remaining speakers are used as imposters. Our cross-validation strategy involved randomizing the batch selection to be used for enrollment and trial. We conducted 5 experiments for each enrollment scenario with different batches and reported the average equal error rate (EER) over these experiments.

Table 1. Phoneme-richness score, speech duration and EER scores associated with select enrollment scenarios

Enrollment scenarios	Speech duration (secs)	P	EER (%)
10 "hello computer" repeats	2.72	0.092	38.49
10 repeats of 4 trigger words	12.66	0.132	26.21
Phoneme-rich passage	11.80	0.152	20.85
1 batch	3.01	0.131	24.79
2 batch	6.03	0.143	20.37
3 batch	9.03	0.150	19.03
4 batch	12.02	0.153	17.88
5 batch	15.06	0.154	17.05

We used an energy-based VAD and GMM-UBM [13] based TI-SR solution in the experiments. Table 1 shows phoneme-richness score, speech duration and EER scores associated with each enrollment scenario for the speaker verification task. Enrollment with 10 repeats of hello computer received the lowest phoneme-richness score and the highest EER. Enrollment with 5 batches of utterances received the highest phoneme-richness score and the lowest EER. Enrollment with the phoneme-rich passage achieves the same EER as 2-3 batches of commands, which is about 11 to 16 unique commands. 10 repeats of 4 trigger words has similar speech duration on the average with the phoneme-rich passage, but EER is significantly lower with the enrollment scenario where a phoneme-rich passage is read.

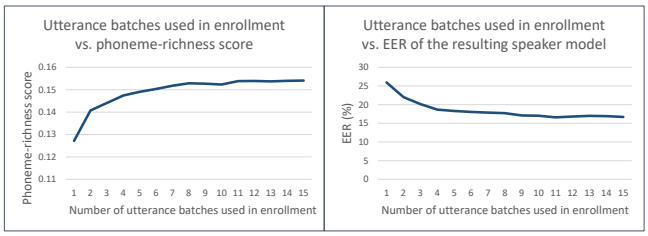

Fig. 4. Phoneme-richness score and EER against the number of utterance batches used for enrollment

Figure 4 shows side by side how phoneme-richness score and EER changes with increasing number of batches used in the enrollment. While phoneme-richness score increases, EER drops from 25% to 15% with the same trend. They both seem to hit a plateau after approximately 11 batches or 60 short commands.

We define success criteria for a sufficiency metric as the correlation between the value of this metric over an utterance pool and the EER when the utterance pool in question is used to create a speaker model. Phoneme-richness metric quality is assessed by Pearsons correlation coefficient between phoneme-richness score and EER on the given enrollment scenarios. This correlation metric measures the linear relationship between two datasets. It varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship. The 18 phoneme-richness score and EER values we measured results in a -0.96 correlation score with p-value 2.21 e-10, indicating a very strong correlation.

One naive way of determining phoneme-richness would be to look at the total speech duration of an utterance pool. This approach is naive because it can be easily fooled to achieve a high sufficiency score by many repeats of the same command. Table 2 shows the comparison of two approaches in terms of correlation with EER. Speech duration metric indicates a much lower correlation with EER compared to phoneme-richness metric. The p-value indicates the probability of incorrectly concluding that a correlation exists. The selection of enrollment scenarios changes the resulting correlation value and therefore correlation values are comparable only when the same enrollment scenarios are used for comparison. In our comparison, we used identical enrollment scenarios.

Table 2. Sufficiency metric comparison using Pearson correlation with EER success criteria

Sufficiency metrics	Correlation with EER	P-value
Phoneme-richness score	-0.96	2.21 e-10
Speech duration	-0.68	0.002

5. SUMMARY

In this paper, we introduced a novel metric to quantify sufficiency of an utterance pool towards training a text-independent speaker model. For achieving fully automated seamless enrollment for text-independent speaker recognition (TI-SR), sufficiency quantification is necessary and no known method exists in the literature for this purpose. Our sufficiency metric, phoneme-richness score, utilizes Gaussian mixture models to model phoneme density for a given language. We use correlation between the value of a sufficiency metric and equal error rate (EER) as success criteria for the sufficiency quantification method. We calculated EER and sufficiency score on a number of enrollment scenarios and found the correlation to be -0.96 for the phoneme-richness metric indicating a highly significant correlation. On the other hand, the naive sufficiency metric speech duration had a -0.68 correlation score.

In addition to making seamless enrollment more robust, sufficiency metrics provide useful information to improve traditional TI-SR enrollment and identification/verification processes. Optimal enrollment durations to achieve target accuracies can be determined using sufficiency metrics. Confidence quantification in multi-session enrollment [3] and model fusion [4, 5] is critical and sufficiency score can be utilized to have an explicit understanding of the quality of the utterances used for enrollment. Sufficiency metrics can also be used during verification/identification for confidence modeling. A higher sufficiency score over a test utterance may indicate a lower expected error rate and therefore higher confidence in the system decision.

As future work, other feature spaces can be explored, although the high correlation between the proposed metric and the TI-SR performance indicates mel-filterbank features contain by far the dominant information for characterizing speakers. Experimenting with low quality data, spontaneous and noisy speech would be necessary to show how the proposed method generalizes for these environmental conditions.

6. REFERENCES

[1] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[2] Jonathan J Huang, "Apparatus and method for voice based user enrollment with video assistance," Aug. 2 2016, US Patent 9,406,295.

[3] Gang Liu, Taufiq Hasan, Hynek Boril, and John HL Hansen, "An investigation on back-end for speaker recognition in multi-session enrollment," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7755–7759.

[4] Niko Brummer, Jan Cernocky, Martin Karafiát, David A van Leeuwen, Pavel Matejka, Petr Schwarz, Albert Strasheim, et al., "Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[5] Ravi P Ramachandran, Kevin R Farrell, Roopashri Ramachandran, and Richard J Mammone, "Speaker recognitiongeneral classifier approaches and data fusion methods," *Pattern Recognition*, vol. 35, no. 12, pp. 2801–2821, 2002.

[6] Alberto Harriero, Daniel Ramos, Joaquin Gonzalez-Rodriguez, and Julian Fierrez, "Analysis of the utility of classical and novel speech quality measures for speaker verification," in *International Conference on Biometrics*. Springer, 2009, pp. 434–442.

[7] Daniel Garcia-Romero, Julian Fiérrez-Aguilar, Joaquin Gonzalez-Rodriguez, and Javier Ortega-Garcia, "On the use of quality measures for text-independent speaker recognition," in *ODYSSEY04-the speaker and language recognition workshop*, 2004.

[8] Daniel Garcia-Romero, Julian Fierrez-Aguilar, Joaquin Gonzalez-Rodriguez, and Javier Ortega-Garcia, "Using quality measures for multilevel speaker recognition," *Computer Speech & Language*, vol. 20, no. 2, pp. 192–209, 2006.

[9] Jonas Richiardi, Krzysztof Kryszczuk, and Andrzej Drygajlo, "Quality measures in unimodal and multimodal biometric verification," in *Signal Processing Conference, 2007 15th European*. IEEE, 2007, pp. 179–183.

[10] Jonas Richiardi and Andrzej Drygajlo, "Evaluation of speech quality measures for the purpose of speaker verification.," in *Odyssey*, 2008, p. 5.

[11] Finnian Kelly, Andrzej Drygajlo, and Naomi Harte, "Speaker verification in score-ageing-quality classification space," *Computer Speech & Language*, vol. 27, no. 5, pp. 1068–1084, 2013.

[12] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[13] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.