LANGUAGE-INDEPENDENT VOICE PASSPHRASE VERIFICATION

Gilles Boulianne

Centre de recherche informatique de Montreal (CRIM), Canada

ABSTRACT

Voice passphrase verification is the task of deciding whether an audio recording contains a given passphrase. It is usually done by evaluating the likelihood of the passphrase reference text given the audio, which requires a different ASR system for each language. Here we look at verification when the passphrase reference is an audio recording instead of a text. We propose a decision likelihood ratio derived from a generative model. Training is unsupervised and needs only audio, without labelling, so the method applies to any language for which recorded audio exists. We report experiments on English and Urdu telephone speech, and show that our model-based likelihood ratio largely outperforms a baseline of DTW based on MFCC feature vectors.

***Index Terms*—** password verification, passphrase verification, dynamic time warping, posteriorgram, unsupervised

1. INTRODUCTION

Voice passphrase verification is used to prevent replay attacks in speaker verification systems: it confirms that a user did repeat the new, unpredictable random passphrase prompted by the system. It can also be used for password reset to verify secret answers known only to the user.

Usually passphrases are specified as text, as in [1]. At verification time, a user utterance is evaluated against a stored text passphrase using an automatic speech recognition (ASR) system. Several components of ASR depend on the language, such as pronunciation rules and acoustic speech models. Their development is costly in terms of effort and time, relies on large transcribed databases and must be repeated for each language. For some languages, the needed resources may not be available so a costly collection effort is needed.

For voice passphrase verification, only audio recordings are used. At verification time, the user utterance is compared to a recorded passphrase spoken by a different speaker. The decision must be robust against factors that make recordings differ even if they contain the same passphrase: individual speaker voice characteristics, variability in speaking rate and pronunciation across speakers, and several other factors such as background noise and transmission channel.

In other applications, speaking rate variations have been handled by time alignment techniques such as Dynamic Time Warping (DTW) [2]. Features like mel-frequency cepstral coefficients (MFCC) have been used in conjunction with DTW to reduce speaker dependency. More recently, combining DTW with features derived from Gaussian models (posteriorgrams) showed promising results for query-by-example [3], unsupervised spoken term detection [4],[5], and mispronunciation detection [6].

Most of these studies have cast their task as a classification or identification problem and use an absolute distance to compare utterances. For verification, [7] trains a model for the distance distribution and bases the decision on the likelihood. [6] trains a Support Vector Machine (SVM) classifier to partition the distance in two classes (accept or reject). Even though these methods may use unsupervised models to obtain a distance between utterances, their decision process requires an additional supervised model or classifier, trained on a corpus labeled with good / bad decisions or, equivalently, a training corpus labeled with the lexical content of each recording.

Here we also use posterior probabilities from a Universal Background Model (UBM) [3][4][5], but we derive a likelihood ratio for the audio password verification decision, based on a probabilistic, generative framework, in a principled way. In contrast to [7] and [6], this likelihood ratio can be compared directly to a threshold for the decision, without the need for an additionnal model or classifier. Since the UBM models used to compute posteriors are also trained without any supervision, training only requires raw audio data, without labels.

In the next sections we present the principle of model-based passphrase verification and the derivation of a likelihood ratio for decision. We will then present experimental results on English and Urdu databases, and compare against baselines of more conventional DTW with MFCC features, and a fully-trained ASR system.

2. MODEL-BASED VOICE PASSPHRASE VERIFICATION

The voice passphrase verification task can be defined formally as follows. Each trial consists of a target passphrase with a unique but unknown lexical content, defined by an audio recording $\mathbf{X}$ from a speaker S_x, and a test recording $\mathbf{Y}$, from another speaker S_y. The system must decide whether $\mathbf{Y}$ is a

This work was funded in part by VoiceTrust and the Ministère de l'économie, innovation et exportation (MEIE) of Gouvernement du Québec.

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

recording of the target passphrase, or equivalently, since the passphrase itself is unknown, decide whether $\mathbf{X}$ and $\mathbf{Y}$ contain the same passphrase.

The first variability to be taken into account comes from differences in timing of the two utterances to be compared. A common approach is to non-linearly map each utterance time to a common time, a process called alignment or warping.

2.1. Dynamic Time Warping

The well-known dynamic time warping algorithm was originally proposed by [2] to find an optimal alignment between two speech utterances. Let $\mathbf{X}$ and $\mathbf{Y}$ be sequences of feature vectors, $\mathbf{x}_n, n = 1, ..., N_x$ and $\mathbf{y}_m, m = 1, ..., N_y$. An alignment path $\tau = \{(n_l, m_l)\}$ is a common warped time axis indexed by $l = 1, ..., L$, which provides for each l a pair of indices (n, m) in the original sequences $\mathbf{X}$ and $\mathbf{Y}$ respectively.

The optimal warping path minimizes a global distance along a particular alignment path τ, over the set T of all possible paths :

$$D(\mathbf{X}, \mathbf{Y}) = \min_{\tau \in T} \frac{1}{L} \sum_{(n,m) \in \tau} d(\mathbf{x}_n, \mathbf{y}_m) \tag{1}$$

2.2. Generative model

Assume $\mathbf{X}$ and $\mathbf{Y}$ are generated by sampling from a collection of models $\mathcal{M} = \{\theta_j\}$, each one generating individual feature vectors $\mathbf{x}_n$ with probability $p(\mathbf{x}_n|\theta_j)$. The generative process is as follows. For each $n = 1, ..., N_x$ we pick a model θ_j from $\mathcal{M}$ with probability $p(\theta_j)$ and generate $\mathbf{x}_n$. Similarly for each $m = 1, .., N_y$ we pick a model from $\mathcal{M}$ and generate $\mathbf{y}_m$. Let's assume that each model in $\mathcal{M}$ represents some sort of lexical subunit. We'll say that $\mathbf{X}$ and $\mathbf{Y}$ have the same lexical content if they are generated by the same sequence of models (null hypothesis H_0). The alternative is that $\mathbf{X}$ and $\mathbf{Y}$ have different lexical content, thus were generated by two unrelated model sequences (alternative hypothesis H_1). The decision is whether H_0 or H_1 is true.

Now consider an alignment path τ. Under the null hypothesis H_0, for each vector pair $\mathbf{x}_n, \mathbf{y}_m$ in the alignment path, the probability that $\mathbf{x}_n$ and $\mathbf{y}_m$ were generated independently, but by the same model θ_j from $\mathcal{M}$ is:

$$\begin{aligned} p(\mathbf{x}_n, \mathbf{y}_m | H_0, \mathcal{M}) &= \sum_j p(\mathbf{x}_n, \mathbf{y}_m | \theta_j) p(\theta_j) \\ &= \sum_j p(\mathbf{x}_n | \theta_j) p(\mathbf{y}_m | \theta_j) p(\theta_j) \end{aligned}$$

(using the conditional independence of $\mathbf{x}_n$ and $\mathbf{y}_m$ given θ_j), so the probability of the complete path when observations are generated from a common sequence of models is:

$$p(\tau | H_0) = \prod_{(n,m) \in \tau} \sum_j p(\mathbf{x}_n | \theta_j) p(\mathbf{y}_m | \theta_j) p(\theta_j) \tag{2}$$

The alternative hypothesis H_1 is that along the alignment path τ, $\mathbf{x}_n$ and $\mathbf{y}_m$ were generated independently from unrelated models:

$$p(\mathbf{x}_n, \mathbf{y}_m | H_1, \mathcal{M}) = \sum_j p(\mathbf{x}_n | \theta_j) p(\theta_j) \sum_i p(\mathbf{y}_m | \theta_i) p(\theta_i)$$

and the probability of the complete path having been generated by two unrelated model sequences is thus:

$$p(\tau | H_1) = \prod_{(n,m) \in \tau} \sum_j p(\mathbf{x}_n | \theta_j) p(\theta_j) \sum_i p(\mathbf{y}_m | \theta_i) p(\theta_i) \tag{3}$$

For a given path τ, we form the likelihood ratio of the null hypothesis H_0 against the alternative hypothesis H_1:

$$\Lambda_\tau(\mathbf{X}, \mathbf{Y} | \mathcal{M}) = \frac{p(\tau | H_0)}{p(\tau | H_1)} \tag{4}$$

Writing (4) in terms of responsibilities $\gamma_{nj} = \frac{p(\mathbf{x}_n|\theta_j)p(\theta_j)}{\sum_j p(\mathbf{x}_n|\theta_j)p(\theta_j)}$ and $\gamma_{mj} = \frac{p(\mathbf{y}_m|\theta_j)p(\theta_j)}{\sum_i p(\mathbf{y}_m|\theta_i)p(\theta_i)}$, and maximizing over all paths, the best likelihood ratio is obtained:

$$\Lambda(\mathbf{X}, \mathbf{Y}) = \max_{\tau \in T} \frac{1}{L} \prod_{(n,m) \in \tau} \sum_j \gamma_{nj} \gamma_{mj} \cdot \frac{1}{p(\theta_j)} \tag{5}$$

We select H_0 whenever $\Lambda(\mathbf{X}, \mathbf{Y})$ is larger than a fixed decision threshold t [1].

The interior summation in (5) is a dot product of two vectors of responsibilities:

$$\sum_j \gamma_{nj} \gamma_{mj} = \boldsymbol{\gamma}_n \cdot \boldsymbol{\gamma}_m \tag{6}$$

where $\boldsymbol{\gamma}_n = \{\gamma_{nj}\}$ and $\boldsymbol{\gamma}_m = \{\gamma_{mj}\}$. Then the negative log of the likelihood ratio of (5) corresponds to the dissimilarity between $\mathbf{X}$ and $\mathbf{Y}$:

$$-\log \Lambda(\mathbf{X}, \mathbf{Y}) = \min_{\tau \in T} \sum_{(n,m) \in \tau} -\log(\boldsymbol{\gamma}_n \cdot \boldsymbol{\gamma}_m) \tag{7}$$

2.3. Gender independent likelihood ratio

For mixed-gender trials, a realistic assumption for most applications is that the gender of reference $\mathbf{X}$ is known (for example, when prompted passphrases are recorded by a known speaker), but the gender of test $\mathbf{Y}$ is not. A well-motivated likelihood ratio for gender independent scoring was proposed in [8] for speaker verification. The idea is to include possible alternatives for gender. In the numerator, the null hypothesis assumes both reference and test to be from the same speaker. In the denominator, all the possible same and cross gender

[1] In practice, experiments show that the term $\frac{1}{p(\theta_j)}$ can be replaced by a constant with negligible effect.

Dataset	Description	N. of speakers	Amount of training	N. of target trials	N. of non-target trials
RsrEng	Singapore English, passphrases, 5 and10 digits sequences	157 m 143 f	8h16 vad 11h33 novad	1600 m 1600 f	640 000 m 640 000 f
PakEng	Pakistan English, 4 digits sequences	223 m 78 f	1h20 vad 4h09 novad	342 m 126 f	133 722 m 17 388 f
PakUrdu	Pakistan Urdu, 4 digits sequences	223 m 78 f	1h15 vad 4h35 novad	512 m 146 f	298 496 m 23 214 f

Table 1. Datasets used. Speaker gender: m (male) or f (female). Voice-activity detection: vad (used) or novad (not used).

combinations, for recordings and models, are taken into account. Using (5) and assuming equiprobable genders for reference and test, the gender-independent likelihood ratio of [8] reduces to:

$$\Lambda_{gi}(\mathbf{X},\mathbf{Y}|\mathcal{M}_M,\mathcal{M}_F) = \frac{1}{2}\Lambda(\mathbf{X}_M,\mathbf{Y}|\mathcal{M}_M) + \frac{1}{2}\Lambda(\mathbf{X}_F,\mathbf{Y}|\mathcal{M}_F) \quad (8)$$

where $\mathbf{X}_M$, $\mathcal{M}_M$ refer to male audio reference and model, and $\mathbf{X}_F$, $\mathcal{M}_F$ refer to female audio reference and model. So a reference audio of each gender is needed and the test passphrase is aligned with each, using the corresponding gender model [2] in (5); the two scores are then combined according to (8).

2.4. Normalization without models

For conventional DTW with MFCC observation vectors, the similarity of (1) and (7) also suggests the following treatment. For the null hypothesis, using $d_e(\mathbf{x}_n,\mathbf{y}_m)$ for the Euclidean distance between $\mathbf{x}_n$ and $\mathbf{y}_m$, we have:

$$D_e(\mathbf{X},\mathbf{Y}|H_0) = \min_{\tau\in T}\frac{1}{L}\sum_{(n,m)\in\tau} d_e(\mathbf{x}_n,\mathbf{y}_m)$$

For the alternative hypothesis, we relax alignment constraints by matching each feature vector in $\mathbf{X}$ against every vector of $\mathbf{Y}$ (and reciprocally), in effect considering $\mathbf{X}$ and $\mathbf{Y}$ as "bag-of-frames". Then we can write, in terms of $d_e(\mathbf{x}_n,\mathbf{y}_m)$:

$$D_e(\mathbf{X},\mathbf{Y}|H_{bof}) = \sum_n \min_m d_e(\mathbf{x}_n,\mathbf{y}_m) + \sum_m \min_n d_e(\mathbf{x}_n,\mathbf{y}_m)$$

And by similarity with log domain equations, the verification score is expressed as a *difference*:

$$D_{diff}(\mathbf{X},\mathbf{Y}) = D_e(\mathbf{X},\mathbf{Y}|H_{bof}) - D_e(\mathbf{X},\mathbf{Y}|H_0) \quad (9)$$

We found that Euclidean distance could not be used without normalization, and using a difference instead of a ratio provided significantly better performance.

3. EXPERIMENTS

We used three datasets originally collected for text-dependent speaker recognition and already divided into training, development and evaluation, without speaker overlap, as detailed in Table 1.

Passphrase verification trials were derived from development sets. Each trial is a test recording by one speaker and either a target recording of same passphrase by a different, random speaker, or a non-target recording of another passphrase by a different, random speaker. Same speaker, different passphrase utterances were excluded, as typical applications use pre-recorded prompts from a non-user.

RsrEng was derived from the RSR2015 database [9]. Passphrases consist of 30 English sentences common to all sessions and speakers, and 5 and 10 digit sequences in random order. PakEng and PakUrdu contain random 4 digit sequences recorded over the mobile phone network, with much more background noise than RsrEng.

MFCC features vectors with 13 static, delta and delta-delta coefficients, for a total of 39 dimensions, were extracted every 10 ms with a 20 ms sliding Hamming window. Cepstral mean of each utterance was subtracted. For voice-activity experiments, we used a self-adaptive VAD [10].

Trials were scored with MFCC features using the normalization of (9), or with UBM posteriors as in (5) and (8). Scores were mapped with a sigmoid to the $[0,1]$ interval and compared against a varying threshold to build a detection error tradeoff (DET) curve.

The amount of data used for training Gaussian mixture UBMs appears in the 4th column of Table 1. After a number of preliminary experiments on each dataset, the number of Gaussians components in UBM models was fixed at 1024, and the smoothing constant to be added to responsibilities at 10^{-12}. Interestingly, those values were optimal for all three datasets despite the varying amounts of data, and were kept for all experiments reported here.

[2] Although not described here, there is a fairly simple procedure to obtain gender dependent UBM models in a completely unsupervised way.

We report development set results with Equal Error Rate (EER), the point on the DET curve where false acceptance rate (FAR) equals false rejection rate (FRR). For RsrEng, a different decision threshold was used within each subset of passphrases, 5-digits and 10-digits, but errors from all trials were aggregated to yield a single EER measurement. When using gender-dependent UBMs, a different threshold was also used according to the UBM gender. Unknown gender results are obtained by adding up cross-gender and same-gender trials, in effect simulating a random choice of reference gender.

3.1. Results

Table 2 compares the use of gender information in scoring for RsrEng. First row simulates an unrealistic situation where test gender is known: a gender-dependent (GD) UBM with same gender as the test can be used. For unknown test gender (2nd row), performance is degraded, but much less for UBM posteriors than for MFCC features, even for GD-UBMs. The last row shows that gender-independent (GI) scoring of (8) improves UBM results, surpassing even known-gender performance of row one. This was observed on all datasets, so all remaining results will be reported with GI scoring (corresponding to the last row of Table 2).

Gender	MFCC	GI-UBM	GD-UBM
Known	3.3%	2.1%	2.1%
Unknown	6.8%	2.7%	2.4%
Independent	3.7%	-	0.8%

Table 2. EER for RsrEng according to gender use in scoring.

Voice-activity detection is investigated in Table 3. The use of VAD reduces EER for both UBMs and MFCCs, with UBMs providing better results. The best scenario corresponds to the last column, where VAD is applied when training UBMs but not at verification time.

Dataset	MFCC		UBM		
	novad	vad	novad	vad	vad-train
PakEng	28.7%	26.5%	18.1%	16.7%	15.2%
PakUrdu	27.3%	25.7%	38.9%	19.9%	14.0%

Table 3. EER when VAD is used in test and training (vad), in training only (vad-train) or not at all (novad). GI scoring.

Although trials were created with a random speaker for each reference, in practice prompted references would be recorded by one ”good” speaker (per gender). To investigate how random references affect results, we selected one ”golden” speaker per gender and dataset, i.e. the one with the best false alarm and rejection statistics. Table 4 shows that UBM scores are less sensitive to the particular choice of speaker, but that it is possible to select a speaker that will provide better results than random references.

Dataset	MFCC		UBM	
	Random	Golden	Random	Golden
PakEng	26.5%	19.4%	15.2%	13.7%
PakUrdu	25.7%	23.0%	14.0%	13.0%

Table 4. EER for with reference audio from random or golden speakers. VAD used in training but not test.

Finally, in Table 5 we compare the unsupervised method proposed here with a more conventional speech recognition approach developed in previous work. The last column (labelled ASR) shows the EER obtained with an in-house GMM-HMM ASR system for English, using text-based references. The ASR system was trained on the RT03 conversational speech training set (2003 NIST Rich Text Transcription), and had 6600 context-dependent triphones for a total of 59K gaussian components. We used the ratio of the forced alignment acoustic likelihood to the more relaxed phoneme recognition acoustic likelihood. The second column (labelled UBM) recalls the best English results from previous tables. For both ASR and UBM, gender-dependent models were used and the same gender-independent scoring method was applied to combine scores. Note that in speaker verification experiments, RsrEng is known to be a clean dataset which yields low error rates [9] while PakEng is more difficult. The table shows that even though relative system performance vary a lot depending on the dataset, the unsupervised UBM-DTW approach proposed here is competitive with a large, fully supervised ASR system.

Dataset	UBM	ASR
RsrEng	0.8%	5.2%
PakEng	13.7%	10.0%

Table 5. EER on English datasets with likelihood ratios from unsupervised UBM (UBM) or fully-supervised speech recognition (ASR).

4. CONCLUSION

We proposed an approach to audio passphrase verification based on DTW with a likelihood ratio using UBM posterior probabilities. UBM training is unsupervised, requiring only audio data, without labelling. We also investigated methods for gender-independent verification, and the impact of voice-activity detection. Together these methods enable deployment of passphrase verification for any language with minimum effort, and with results competitive with a language-specific, ASR based system.

5. REFERENCES

[1] D. Yu, Y.C. Ju, and A. Acero, "An effective and efficient utterance verification technology using word n-gram filler models," in *Proc. Interspeech*, 2006, pp. 2318–2321.

[2] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE TASSP*, vol. 26, no. 1, pp. 43–49, 1978.

[3] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Proc. ASRU*, 2009, pp. 421–426.

[4] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *Proc. ASRU*, 2009, pp. 398–403.

[5] Y. Zhang and J. R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *Proc. ICASSP*, 2010, pp. 4366–4369.

[6] A Lee and J. R. Glass, "A comparison-based approach to mispronunciation detection," in *Proc. SLT*, 2012, pp. 382–387.

[7] Q. Liu, Z.-Q. Huang, Y.-B. Hou, and R. Chen, "Utterance verification On DTW based speech recognition using likelihood," in *ICCASM*, 2010, pp. 427–431.

[8] M. Senoussaoui, Patrick Kenny, Niko Brummer, Edward de Villiers, and Pierre Dumouchel, "Mixture of PLDA Models in i-vector Space for Gender-Independent Speaker Recognition.," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 25–28.

[9] A. Larcher, K. A. Lee, Bin Ma, and H. Li, "RSR2015: Database for Text-Dependent Speaker Verification using Multiple Pass-Phrases.," in *Proc. Interspeech*, Portland, Oregon, USA, Sept. 2012.

[10] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *Proc. ICASSP*, 2013, pp. 7229–7233.