SYNTHETIC SPEECH DETECTION BASED ON SELECTED WORD DISCRIMINATORS

Phillip L. De Leon and Bryan Stewart

New Mexico State University
Klipsch School of Electrical and Computer Engineering
Las Cruces, New Mexico, U.S.A.

ABSTRACT

Speaker verification (SV) systems have been shown to be vulnerable to imposture using speech synthesizers. In this paper, we extend previous work in detecting synthetic speech by analyzing words which provide strong discrimination between human and synthetic speech. The research is applicable to authentication systems based on text-dependent SV where the user is prompted to speak a certain utterance which can be chosen by the designer. Our results show that this approach to synthetic speech detection leads to higher accuracies than other proposed approaches. Using various corpora to train and test, our results show 98% accuracy in correctly classifying both human and synthetic speech.

Index Terms— speaker recognition, speech synthesis, security

1. INTRODUCTION

A speaker verification (SV) system accepts or rejects a claimed identity based on a voice sample from the speaker. SV systems have been shown to be vulnerable to speech synthesis [1], voice conversion [2], and voice mimicking [3] by accepting an identity claim based on a manipulated or synthesized speech signal.

In [4], the authors extended their prior research in [5] by removing the dependency of training the classifier with a synthetic voice matched to each human enrolled in the system. This was accomplished by using transcoded human speech during training as a surrogate to the synthetic counterpart. In [4], the SV and synthetic speech detection (SSD) system were trained and tested on $\approx$90s and $\approx$30s speech signals, respectively, whereas the synthetic speech classifier in [5] was trained on only 10s of voiced speech per speaker. The SSD in [4] reported 100% accuracy in classifying both human and synthetic speech when trained on synthetic voices matched to each human enrolled in the system. The SSD systems had 100% accuracy in classifying human speech and 90% accuracy in classifying synthetic speech when trained with the surrogate synthetic speech; the SSD also had an equal error rate (EER) of 97%. Though these results are improvements to previously reported accuracies, the SSD features are entirely phase-based with an assumed knowledge of the phase model used to create the synthetic speech. The results previously stated for "coded speech" were generated using a vocoder with a minimum phase model. For example, when a vocoder with different phase characteristics is used to create the surrogate synthetic speech dataset and therefore train the SSD, the authors report that the accuracy fell from 90% down to 6.3% for accurately classifying synthetic speech [4].

In [6], we proposed a classifier based on features extracted from image analysis of pitch patterns that did not require synthetic models matched to humans in the SV system or any *a priori* information regarding speech synthesizers. These features which include the mean pitch stability, mean pitch stability range, and jitter were found to provide good discrimination between human and synthetic speech. The classifier modeled the distribution of synthetic speech feature vectors as a multivariate Gaussian distribution with a diagonal covariance matrix. A decision threshold was then set by computing the likelihoods of the training feature vectors and adjusting for combined highest accuracy. We trained the classifier based on features extracted from human speech (NIST2002 corpus) and synthetic speech (2008 and 2011 Blizzard Challenge along with Festival pre-built voices). The classifier was evaluated using speech from the Switchboard corpus, Resource Management corpus, and synthetic speech generated from Festival trained on the Wall Street Journal (WSJ) corpus. Classification of human, synthetic speech was 98%, 96% accurate, respectively.

The motivation for this work is based on informal listening tests where it was observed that certain words *sound* more synthetic than others regardless of the synthesizer or vocoder. Although these sounds are likely rooted in unnatural-modeling of certain phonemes, our study begins by analyzing common words in our corpora which could serve as the basis for improved discrimination. The target application is authentication systems based on *text-dependent* SV where the user is prompted to speak a certain utterance which can be chosen by the designer. In this case, the utterance can contain many of these discriminating words thus improving the accuracy of synthetic speech detection. Unlike [6], this paper 1) leverages sub-utterance information i.e. word segments of the claimant's utterance, 2) statistically models each

word's features using a multivariate Gaussian distribution, and 3) utilizes a maximum likelihood (ML) classifier with a weighted mean feature vector based on the Bhattacharyya distance measure.

This paper is organized as follows. In Section 2, we review the pitch pattern features proposed in [6,7]. In Section 3, we discuss the analysis of discriminating words and propose a ML classifier using a Bhattacharyya weighted mean feature vector. In Section 4, we describe the corpora used in training and testing and provide classifier results. Future work is discussed in Section 5. Finally, in Section 6, we conclude the paper.

2. IMAGE-BASED PITCH PATTERN FEATURES

In this section, we briefly summarize the pitch pattern features first proposed in [7] and the image-based pitch pattern features recently proposed in [6].

2.1. Pitch Pattern

The pitch pattern, $\phi(t,\tau)$, is calculated by dividing the short-range autocorrelation function, $r(t,\tau)$ by a normalization function, $p(t,\tau)$ which is proportional to the frame energy [7]

$$\phi(t,\tau) \quad = \quad \frac{r(t,\tau)}{p(t,\tau)}. \tag{1}$$

Once the pitch pattern is computed, we segment into a binary pitch pattern image through the rule

$$\phi_{\text{seg}}(t,\tau) \quad = \quad \begin{cases} 1, & \phi(t,\tau) \geq \theta_t \\ 0, & \phi(t,\tau) < \theta_t \end{cases} \tag{2}$$

where θ_t is a threshold set to half the pitch pattern peak value at time t. In this paper, we compute $\phi(t,\tau)$ for $2 \leq \tau \leq 20$ms and set $\theta_t = 1/\sqrt{2}$ for all t. An example pitch pattern image is shown in Fig. 1.

2.2. Image Analysis of the Pitch Pattern

Extracting features from the pitch pattern is a multi-step process and includes 1) silence removal, 2) voiced / unvoiced segmentation, 3) computation of the pitch pattern, and 4) image analysis. In the fourth step, image processing of the segmented binary pitch pattern is performed in order to extract the connected components, i.e. black regions in Fig. 1. This processing includes determining the bounding box and area of a connected component which are then used to filter out very small and irregularly-shaped components. The resulting connected components are then analyzed and used to compute mean pitch stability, μ_S; mean time stability bandwidth, μ_B; and jitter, J which are defined next. The feature vector used in classification is given by

$$\mathbf{x} \quad = \quad [\mu_S, \mu_R, J]\,. \tag{3}$$

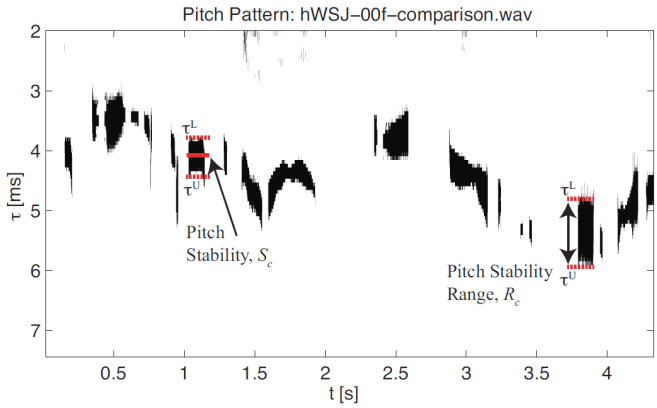

Fig. 1. Segmented binary pitch pattern image from a human speech signal [6]. The phrase is "The female produces a litter of two to four young in November." Pitch stability S_c, pitch stability range R_c, upper edge τ^{U}, and lower edge τ^{L} are denoted.

2.3. Mean Pitch Stability

The *pitch stability* of connected component, c is the average value of τ over the connected component

$$S_c \quad = \quad \frac{1}{T}\int_c \left[\frac{\tau^{\text{U}}(t)+\tau^{\text{L}}(t)}{2}\right] dt \tag{4}$$

where T is the time-support of c and where U and L denote the upper and lower edges of τ, respectively (see Fig. 1) [6]. The mean pitch stability is calculated as

$$\mu_S \quad = \quad \frac{1}{C}\sum_{c=1}^{C} S_c \tag{5}$$

where C is the number of connected components in the speech signal.

2.4. Mean Pitch Stability Range

The *pitch stability range* of connected component, c is the average range of τ over the connected component

$$R_c \quad = \quad \frac{1}{T}\int_c \left[\tau^{\text{U}}(t)-\tau^{\text{L}}(t)\right] dt \tag{6}$$

(see Fig. 1) [6]. The mean pitch stability range is calculated as

$$\mu_R \quad = \quad \frac{1}{C}\sum_{c=1}^{C} R_c. \tag{7}$$

2.5. Jitter

The pitch pattern jitter, J is computed as follows. The peak lag for connected component, c at time t is calculated as

$$\phi_c'(t) \quad = \quad \max_{\tau} \phi(t,\tau) \tag{8}$$

and the variance of the peak lags for connected component, c is calculated as

$$\sigma_c^2 \quad = \quad \mathrm{var}\left[\phi_c'(t)\right]. \tag{9}$$

The pitch pattern jitter, J is then the average of the peak lag variances of the connected components [6]

$$J \quad = \quad \frac{1}{C}\sum_{c=1}^{C}\sigma_c^2. \tag{10}$$

2.6. Segmental vs. Supra-segmental Features

Vocal tract features, such as those based on mel-frequency cepstral coefficients (MFCCs), are normally *segmental* and based on short-time frames. As shown in [1, 4], MFCCs are insufficient in discriminating between synthetic and natural speech. On the other hand, connected components extracted from the binary pitch pattern image are *supra-segmental* features extracted across many frames. It is our hypothesis that the co-articulation, or supra-segmental characteristics of the pitch pattern for synthetic speech, differs from that of human speech and to a greater extent in certain words. To illustrate this point, Fig. 2, shows a scatter plot of feature vectors for the top 20 human and synthetic word models with the largest separation as calculated by the Bhattacharyya distance measure.[1] It is evident that for human speech, these features lie in a compact and distinct space as compared to synthetic speech. Selection of these words is described in the next section.

3. MAXIMUM LIKELIHOOD CLASSIFIER

In [6], the classification of speech as human or synthetic was based on an averaged connected component feature vector extracted from the utterance. We propose a ML classifier based on the log-likelihoods computed from the weighted mean feature vector extracted at the *word-level*. During training, we model the distribution of human and synthetic speech feature vectors as multivariate Gaussian distributions with diagonal covariance matrices, $\mathcal{N}^{\mathrm{hum}}(\boldsymbol{\mu}_{\mathrm{hum}}, \boldsymbol{\Sigma}_{\mathrm{hum}})$ and $\mathcal{N}^{\mathrm{syn}}(\boldsymbol{\mu}_{\mathrm{syn}}, \boldsymbol{\Sigma}_{\mathrm{syn}})$, respectively. Also during training, each unique word's feature vectors (collected from all human or all synthetic training speech), are individually modeled using Gaussian distributions, $\mathcal{N}_n^{\mathrm{hum}}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ and $\mathcal{N}_n^{\mathrm{syn}}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$

[1]The selected words are: "about", "and", "be", "boy", "but", "down", "if", "look", "many", "more", "most", "much", "my", "no", "nothing", "so", "the", "them", "time", and "with."

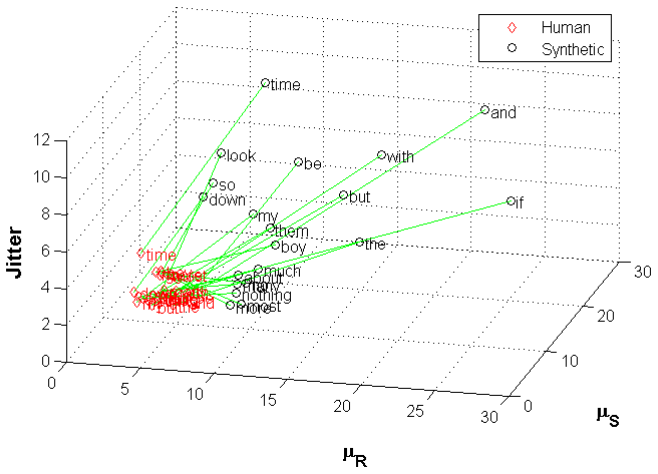

Fig. 2. Scatter plot of the mean pitch stability, μ_S; mean pitch stability range, μ_R; and jitter, J, of 20 word models with the largest separation as calculated by Bhattacharyya distance measure. Human speech features lie in a compact and distinct space as compared to synthetic speech features.

where n denotes the word index. A distance measure (described below) is also computed between the human and synthetic word models. The unique words and distance measures are then stored in a lookup table indexed by n.

In the test stage, $\mathbf{x}_n$ is the feature vector extracted from the nth corresponding word and the weighted mean feature vector is given by

$$\mathbf{x} = \sum_{n=1}^{N} D_n \mathbf{x_n} \tag{11}$$

where D_n is a distance measure between $\mathcal{N}_n^{\mathrm{hum}}$ and $\mathcal{N}_n^{\mathrm{syn}}$ and N is the number of words in the test utterance. The log-likelihood ratio is then given by

$$\Lambda = \log p(\mathbf{x}|\mathcal{N}^{\mathrm{hum}}) - \log p(\mathbf{x}|\mathcal{N}^{\mathrm{syn}}) \tag{12}$$

and the utterance is determined to be human if

$$\Lambda \geq \theta \tag{13}$$

where θ is the decision threshold.

There are many distance measures that could be used to calculate D_n in (11) [8], however, we have found the Bhattacharyya distance measure works well. The Bhattacharyya distance between Gaussian pdfs, $\mathcal{N}_i(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and $\mathcal{N}_j(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ is given by [9]

$$\begin{aligned} D_{\mathrm{B}}(\mathcal{N}_i \| \mathcal{N}_j) \quad &= \quad \frac{1}{8}\left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\right)^{\top}\left[\frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2}\right]^{-1}\left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\right) \\ &\quad + \frac{1}{2}\ln\left(\frac{\left|\left(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j\right)/2\right|}{\sqrt{\left|\boldsymbol{\Sigma}_i\right|\left|\boldsymbol{\Sigma}_j\right|}}\right) \end{aligned} \tag{14}$$

where $|\cdot|$ denotes matrix determinant. The first term on the right side of the equality, measures separation due to the Gaussian pdfs' mean vectors, $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$, while the second term measures separation due to the Gaussian pdfs' covariance matrices, $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Sigma}_j$.

The distance between each human word model and corresponding synthetic word model is calculated during training. During testing D_n is normalized so that

$$\sum_{n=1}^{N} D_n(\mathcal{N}_n^{\text{hum}} \| \mathcal{N}_n^{\text{syn}}) \quad = \quad 1. \tag{15}$$

The weights used in the classifier emphasize the feature vectors of the word models that exhibit greater separability between human and synthetic speech. Conversely, the weights de-emphasize the feature vectors of word models that are similar.

4. EXPERIMENTS AND RESULTS

As part of this research, we collected synthetic speech material from a variety of sources as well as directly synthesized speech. The Festival Speech Synthesis System v2.1 was used to synthesize the speech signals used during this research. The WSJ corpus was used to construct 283 different speaker models using a speaker-adaptive, HMM-based speech synthesis system, H Triple S (HTS). These WSJ HTS speaker models were used in Festival to generate the synthetic WSJ speech. Resource Management (RM) voices were obtained from the "Voices of the World" (VoW) demonstration system hosted at The Centre for Speech Technology Research [10]. RM speaker models were generated using a speaker-adaptive HTS similar to the WSJ speaker models [11].

The human TIMIT corpus [12] has 630 speakers, 296 phonetically-balanced sentences, and a total of 4893 unique words. Each WSJ voice model was used to synthesize all 296 phonetically-balanced TIMIT sentences resulting in 283 synthetic speakers each uttering 4893 unique words. The human Switchboard-1 [13] corpus was separated into word segments according to the 500 word vocabulary defined in [14]. The synthetic RM corpus is comprised of 157 synthesized voices each uttering 106 unique words. In this research, there were a 106 common unique words that were spoken by each speaker of the four corpora. Half of the human speakers and half of the synthetic speakers were used for training. The other half of the human and synthetic speakers were used for testing the classifier. We chose 221 human speakers at random from the available corpora in order to match the number of synthetic speakers used in testing; there were no speakers in common between the training and testing datasets. Speech corpora usage is summarized in Table 1.

The results presented in [6] used a likelihood classifier and a different set of training and testing corpora than presented in this paper. For this paper, the classifier used in [6] was re-evaluated with the training and testing corpora presented here (Table 1) and the results showed 96%, 92% classification accuracy for human, synthetic speech, respectively. The proposed classifier results in improved classification accuracy of 98%, 98% for human, synthetic speech, respectively.

The results for the proposed ML classifier using the Bhattacharyya weighted mean feature vector are not only better than those presented in [5] but furthermore, do not require development of a synthetic voice matched to each human enrolled in the system. Furthermore, the results are better than [4] without assuming knowledge of the phase model used to create the synthetic speech features or any other prior information regarding the synthesizer. In addition, greater classification accuracy was achieved by the proposed classifier, compared to recent work in [6], by leveraging sub-utterance information of the claimant's speech signal.

Table 1. Speech corpora and number of speakers used for training and testing the proposed weighted, ML classifier. Speakers are unique to either training or testing datasets.

	Training (Num. of Spkrs.)	Testing (Num. of Spkrs.)
Human	TIMIT (315) Switchboard (259)	TIMIT (111) Switchboard (110)
Synthetic	WSJ (141) RM (78)	WSJ (142) RM (79)

5. FUTURE WORK

Certain words do provide stronger discrimination between human and synthetic speech. However, modeling a large number of words may be impractical. Our future work includes modeling feature vectors at the phoneme-level where we have observed large separation distances in the feature vectors for certain phonemes. It is anticipated that generative and discriminative classifiers operating at the phoneme-level will results in increased accuracy.

6. CONCLUSIONS

In this paper, we proposed a maximum likelihood classifier using a Bhattacharyya weighted mean feature vector based on the words in a speaker's utterance. We used the synthetic WSJ speech, synthetic RM voices obtained from [10], human speech from the TIMIT corpus, and human speech from the Switchboard-1 corpus. Results show 98% accuracy in correctly classifying human speech and 98% accuracy in correctly classifying synthetic speech. The classifier presented here provided greater discrimination between human and synthetic speech compared to the recent research results.

7. REFERENCES

[1] P. L. De Leon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*, 2010, pp. 151–158.

[2] T. Kinnunen, Z. Wu, K. Lee, F. Sedlak, E. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Kyoto, Japan, 2012, pp. 4401–4404.

[3] Y. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proc. Int. Symp. Intelligent Multimedia, Video and Speech Process.*, Oct. 2004, pp. 145 – 148.

[4] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of hmm-based synthetic speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 8, pp. 2280–2290, Oct 2012.

[5] P.L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 4844–4847.

[6] Phillip L. De Leon, Bryan Stewart, and Junichi Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *Proc. Interspeech*, Portland, Oregon, USA, Sept 2012.

[7] A. Ogihara, H. Unno, and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE Trans. Fundamentals*, vol. E88, no. 1, pp. 280–286, Jan. 2005.

[8] S. Cha, "Comprehensive survey on distance/simialrity measures between probability density functions," *Int. J. Mathematical Models and Methods in Applied Sciences*, vol. 1, pp. 300–307, Sept 2007.

[9] B. Mak and E. Barnard, "Phone clustering using the bhattacharyya distance," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. IEEE, 1996, vol. 4, pp. 2005–2008.

[10] "`http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/map-new.html`," 2010.

[11] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Speech, Audio, and Language Process.*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[12] J.S. Garofolo, *TIMIT: Acoustic-phonetic Continuous Speech Corpus*, Linguistic Data Consortium, 1993.

[13] J.J. Godfrey and E. Holliman, "Switchboard-1 release 2," *Linguistic Data Consortium, Philadelphia*, 1997.

[14] Simon King, Chris Bartels, and Jeff Bilmes, *SVitchboard 1: Small Vocabulary Tasks from Switchboard 1*, pp. 3385–3388, International Speech Communication Association, 2005, ISSN: 1990-9772.