Synthetic Speech Discrimination using Pitch Pattern Statistics Derived from Image Analysis

Phillip L. De Leon[1]*, Bryan Stewart*[1]*, and Junichi Yamagishi*[2]

[1]New Mexico State University, Klipsch School of Elect. and Comp. Eng., Las Cruces, N.M., U.S.A.
[2]University of Edinburgh, Centre for Speech Technology Research (CSTR), Edinburgh, U.K.
{pdeleon, brystewa}@nmsu.edu, jyamagis@inf.ed.ac.uk

Abstract

In this paper, we extend the work by Ogihara, et al. to discriminate between human and synthetic speech using features based on pitch patterns. As previously demonstrated, significant differences in pitch patterns between human and synthetic speech can be leveraged to classify speech as being human or synthetic in origin. We propose using mean pitch stability, mean pitch stability range, and jitter as features extracted after image analysis of pitch patterns. We have observed that for synthetic speech, these features lie in a small and distinct space as compared to human speech and have modeled them with a multivariate Gaussian distribution. Our classifier is trained using synthetic speech collected from the 2008 and 2011 Blizzard Challenge along with Festival pre-built voices and human speech from the NIST2002 corpus. We evaluate the classifier on a much larger corpus than previously studied using human speech from the Switchboard corpus, synthetic speech from the Resource Management corpus, and synthetic speech generated from Festival trained on the Wall Street Journal corpus. Results show 98% accuracy in correctly classifying human speech and 96% accuracy in correctly classifying synthetic speech.

Index Terms: Speaker recognition, Speech synthesis, Security

1. Introduction

State-of-the-art text-to-speech (TTS) systems are capable of generating high-quality, natural sounding speech using small amounts of non-ideal speech data from a targeted person [1], [2]. These systems therefore may pose a risk in speaker recognition (SR) systems. In particular, system access through voice authentication may be vulnerable through attacks using speech synthesizers. Prior research into the problem of imposture from synthetic speech and vulnerability of SR systems to synthetic speech can be found in [3–6]

In [7], Ogihara, et al. proposed to discriminate between human and synthetic speech using features extracted from the pitch pattern[1]. The pitch pattern is calculated as a normalized, short-range, auto-correlation of a speech signal over a 2-20ms range. The authors compared the use of a speaker's time stability and pitch pattern peak, lower half, upper half, and half bandwidth as features to discriminate between human and synthetic speech. The research used 100 samples of human speech from a male subject and generated the synthetic speech using the method proposed in [3]. Decision thresholds based on time stability and pitch pattern measures were obtained from 20 human and 20 synthetic speech samples out of the 100. This process was performed on a total of five individuals with the "half bandwidth" providing the best average performance of correct rejection of synthetic speech, ranging from 93.3% to 100%.

In [6], the relative phase shift (RPS) of voiced speech was used to discriminate between human and synthetic speech for a speaker verification (SV) application. In this work, the Linguistic Data Consortium (LDC) Wall Street Journal (WSJ) corpus (283 speakers) was used for human speech and synthetic voices were constructed for each WSJ speaker. RPS-based feature vectors extracted from both human and synthetic speech were then used to train a Gaussian Mixture Model (GMM) and classification was based on maximum likelihood (ML). The results using the WSJ corpus were 88% of the synthetic speech was classified correctly and 4.2% of the human speech was classified incorrectly. Although the work used a more sophisticated TTS and a much larger corpus than [7], training the classifier required development of a synthetic voice matched to each human enrolled in the system which is not practical.

In this paper, we also seek to develop a system which can accurately classify whether speech is human or synthetic. Unlike [6], however, we wish to train our system using any available synthetic speech without regard to whether it is matched to the human speech used to train the classifier. Thus our system aims to build a more general synthetic speech detection model without restrictions on the training data other than we have a reasonably large number of human speech signal examples and a reasonably large number of synthetic speech signal examples. Our approach in this paper extends the work in [7] by 1) using a novel image processing approach to extract features based on statistical measures of the pitch pattern, 2) proposing an additional feature based on jitter, 3) utilizing a classifier based on a multivariate Gaussian model of the feature distributions, and 4) evaluating the system using a much larger evaluation corpus.

This paper is organized as follows. In Section 2, we review the pitch pattern calculation proposed in [7] and in Section 3 describe the pitch pattern features we use in the classifier. In Section 4, we describe the classifier and various corpora used in training and testing and provide results. In Section 5, we discuss our future research and in Section 6, we conclude the paper.

2. Pitch Pattern

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

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

[1]In [7] as well as this paper, pitch is used synonomously with fundamental frequency, F_0.

The short range auto-correlation function is given by

$$r(t,\tau) \quad = \quad \int_{-\tau/2}^{\tau/2} x(t+\xi-\tau/2)x(t+\xi+\tau/2)\,d\xi \quad (2)$$

and is similar to the short-time autocorrelation function for multiple lag inputs. The normalization function

$$p(t,\tau) \quad = \quad \frac{1}{2}\int_{-\tau/2}^{\tau/2} x^2(t+\xi-\tau/2)\,d\xi + \frac{1}{2}\int_{-\tau/2}^{\tau/2} x^2(t+\xi+\tau/2)\,d\xi \quad (3)$$

is proportional to the frame energy [7].

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

$$\phi_{\text{seg}}(t,\tau) \quad = \quad \begin{cases} 1, & \phi(t,\tau) \geq \theta_t \\ 0, & \phi(t,\tau) < \theta_t \end{cases} \quad (4)$$

where θ_t is a threshold set to half the pitch pattern peak value at time t. An example pitch pattern image is shown in Fig. 1. 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.

3. Feature Extraction from Pitch Pattern

Extracting useful features from the pitch pattern is a multi-step process illustrated in Fig. 2 and includes 1) silence removal, 2) voiced/unvoiced segmentation, 3) computation of the pitch pattern, and 4) image analysis. First, silence is removed from the speech signal using an adaptive voice activity detector (VAD) [8]. Second, the resulting signal is segmented into voiced and unvoiced speech using a frame-based zero crossing and energy detector (20ms frames) which is illustrated in Fig. **??** [9]. Third, the pitch pattern from voiced speech segments is computed using (1) and segmented using (4) to form a binary image.

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. We have observed that very small and irregularly-shaped connected components are artifacts of the speech signal and not useful in feature extraction. The resulting connected components are then analyzed and used to compute the following statistics-based features (defined below): mean pitch stability, μ_S; mean time stability bandwidth, μ_B; and jitter, J. Our proposed image processing-based approach, which determines parameters on a per-connected component basis and then computes statistics over the connected components of the utterance, is in contrast to the features used in [7].

3.1. Mean Pitch Stability

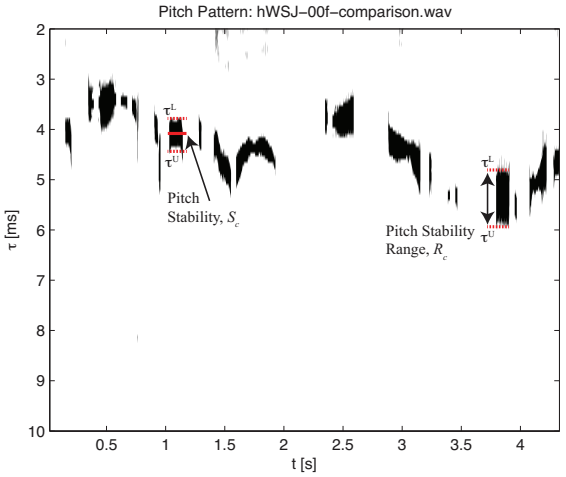
The *pitch stability* of connected component, c is the average value of τ over the connected component

$$S_c \quad = \quad \frac{1}{T}\int_c \left[\frac{\tau^{\text{U}}(t)+\tau^{\text{L}}(t)}{2}\right] dt \quad (5)$$

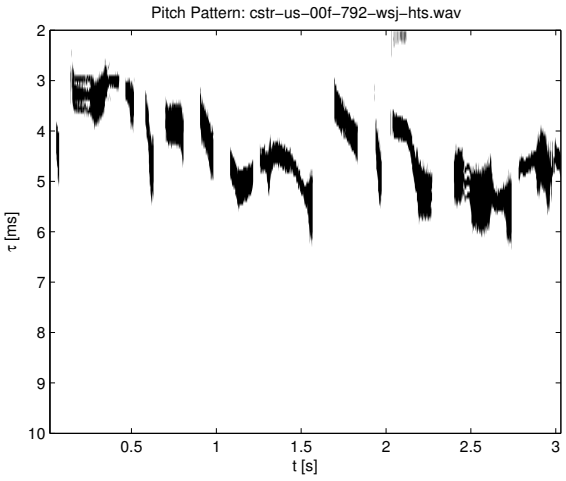
where T is the time-support of c and where U and L denote the upper and lower edges of τ, respectively [see Fig. 1(a)]. The mean pitch stability is calculated as

$$\mu_S \quad = \quad \frac{1}{C}\sum_{c=1}^{C} S_c \quad (6)$$

where C is the number of connected components in the speech signal. In [7], the authors compute a time stability feature (we prefer calling this "pitch stability") through a more complex process involving several thresholding procedures.



(a)



(b)

Figure 1: Segmented binary pitch pattern image from (a) human speech signal and (b) synthetic speech signal. In both plots 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 in (a).

3.2. Mean Pitch Stability Range

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

$$R_c \quad = \quad \frac{1}{T}\int_c \left[\tau^{\text{U}}(t)-\tau^{\text{L}}(t)\right] dt \quad (7)$$

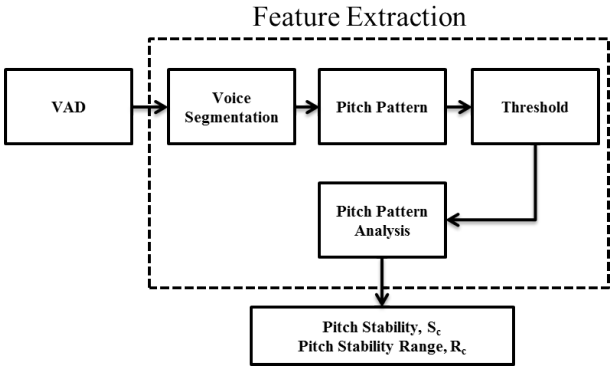


Figure 2: Feature Extraction Diagram.

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

$$\mu_R = \frac{1}{C}\sum_{c=1}^{C} R_c. \quad (8)$$

In [7], the authors compute time stability bandwidth (we prefer calling this "pitch stability range") through a more complex process dependent on peak values in the pitch pattern whereas our process is simplified by setting $\theta_t = 1/\sqrt{2}$ for all t in (4).

3.3. Jitter

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

$$\phi'_c(t) = \max_{\tau} \phi(t,\tau) \quad (9)$$

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

$$\sigma_c^2 = \text{var}\left[\phi'_c(t)\right]. \quad (10)$$

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

$$J = \frac{1}{C}\sum_{c=1}^{C} \sigma_c^2. \quad (11)$$

3.4. Comments

In summary, for the voiced segments of a speech signal the segmented binary pitch pattern is computed with (4); image analysis is performed as described in the fourth step; and mean pitch stability (6), mean pitch stability range (8), and jitter (11) are computed and used to form the feature vector

$$\mathbf{x} = [\mu_S, \mu_R, J]. \quad (12)$$

Based on informal listening tests, state-of-the-art synthetic speech is often hyperarticulated which usually correlates to a larger time stability bandwidth. In addition, because it is difficult to precisely model human physiological features required to properly synthesize natural speech, we hypothesize that synthetic speech will also have a different mean pitch stability than human speech. Finally, we have observed that co-articulation, the transition from one phoneme to the next, of synthetic speech occurs more rapidly than in human speech where co-articulation is smooth and relaxed and this difference is captured by the jitter of the pitch pattern.

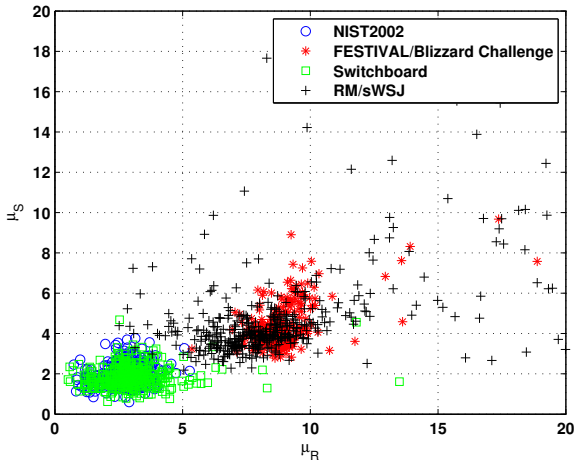


Figure 3: Scatter plot of the mean pitch stability, μ_S and mean pitch stability range, μ_R from speech used in training and evaluation. Human speech features [NIST2002 (blue circle) and Switchboard (green square)] lie in a compact and distinct space as compared to synthetic speech features [Blizzard Challenge (red asterisk) and Resource Management (RM)/synthetic Wall Street Journal (WSJ) (black plus)].

Vocal tract features, such as MFCC, have been insufficient in discriminating between synthetic and natural speech [10]. Vocal tract features are normally segmental (frame-level), multi-dimensional features. On the other hand, the pitch pattern is a scalar time-series sequence–a supra-segmental, long-span feature across many frames. It is our hypothesis that the co-articulation or supra-segmental characteristics of the pitch pattern for synthetic speech, may differ from that of natural speech.

In Fig. 3, we show a scatter plot of the mean pitch stability, μ_S and mean pitch stability range, μ_R from speakers in corpora used in the training and evaluation. It is evident that for human speech (blue circles and green squares), these features lie in a compact and distinct space as compared to synthetic speech (red asterisks and black plus).

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 speech from 15 speaker models included in the system which are based on a diphone synthesizer [11]. Blizzard Challenge voices (total of 226), from the 2008 and 2011 competitions [11–13], were obtained from [14]. We used the WSJ corpus 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 [15]. RM speaker models were generated using a speaker-adaptive HTS similar to the WSJ speaker models [1].

For the synthetic speech used in training the classifier, we used the pre-built Festival voice models to synthesize the ten standard, phonetically-balanced TIMIT sentences beginning with, "She had your dark suit in greasy wash water all year...." This resulted in 15 synthetic speech signals that are 15-30s in duration. The Blizzard Challenge synthetic speech utterances were limited to the first 30s of speech and resulted in 152 and 59 speech signals from the 2008 and 2011, respectively competitions. For the human speech used in training the classifier, we used the NIST2002 corpus (total of 330 speakers) with each signal approximately 30s in length.

We evaluated the classifier using human speech from the Switchboard corpus (352 speakers) and synthetic speech (518 synthesized voices) from the synthetic WSJ voices [2], [5] and the synthetic RM voices, as noted above [15]. The synthetic WSJ voices were generated using the TIMIT sentences and the synthetic RM voices uttering, "Finally a little girl did come along and she was carrying a basket of food." Speech corpora usage is summarized in Table 1.

Table 1: Speech corpora used for training and testing the classifier

	Training	Testing
Human	NIST2002	Switchboard
Synthetic	Blizzard 2008/2011, Festival pre-built voices	WSJ, RM

Feature vectors in (12) are extracted from human and synthetic training speech. The distribution of synthetic speech feature vectors is modeled as a multivariate Gaussian distribution with a diagonal covariance matrix. A decision threshold is then set by computing the likelihoods of the training feature vectors and adjusting for combined highest accuracy. Using the test speech, results show classification accuracy of 98% for human speech and 96% for synthetic speech. The results for classification of synthetic speech are better than those presented in [6] but without the complication of requiring development of a synthetic voice matched to each human enrolled in the system. In addition, our results are as good or better than [7] but using a much larger evaluation set and a classifier trained with a corpus different than that used in testing.

5. Future Work

The feature vector in (12) is extracted from the pitch pattern for voiced segments within an utterance. These vectors are then collectively modeled as a Gaussian distribution. Our future work includes modelling the vectors at the phoneme-level where we have observed large separation distances in the feature vectors for certain phonemes. Classifiers based at the phoneme-level could result in increased accuracy.

6. Conclusions

In this paper, we have used mean pitch stability, mean pitch stability range, and jitter as features extracted from image analysis of pitch patterns for discrimination between human and synthetic speech. We have developed a classifier based on a Gaussian distribution of these features. We used pre-built Festival voices, obtained synthetic voices from the Blizzard Challenge and human speech from the NIST2002 corpus for training the classifier. The classifier was evaluated using human speech from the Switchboard corpus and synthetic WSJ and RM voices. Results show 98% accuracy in correctly classifying human speech and 96% accuracy in correctly classifying synthetic speech.

7. References

[1] 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. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[2] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis – Analysis and application of TTS systems built on various ASR corpora," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 5, pp. 984–1004, Jul 2010.

[3] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture against a speaker verification system using synthetic speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. J83-D-II, no. 11, pp. 2283–2290, Nov 2000.

[4] P. L. De Leon, V. R. Apsingekar, M. Pucher, and J. Yamagishi, "Revisiting the security of speaker verification systems against imposture using synthetic speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, USA, 2010.

[5] P. L. De Leonn, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Proc. IEEE Speaker and Language Recognition Workshop (Odyssey)*, Brno, Czech Republic, 2010.

[6] P. L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, USA, 2011.

[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. Kuo, B. Lee, and W. Tian, *Real-Time Digital Signal Processing,Implementations and Applications, 2nd edition*. Wiley, 2006.

[9] R. Cachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," *ASEE*, Dec. 2008. [Online]. Available: {http://www.asee.org/documents/zones/zone1/2008/student/ASEE12008_0044_paper.pdf}

[10] P. L. De Leon, M. Pucher, and J. Yamagishi, "In review: Evaluation of speaker verification security and detection of hmm-based synthetic speech," in *IEEE Trans. Audio, Speech, Language Processing*, Dallas, USA, 2012.

[11] A. W. Black, P. Taylor, and C. Richard, "The Festival Speech Synthesis System," 1997. [Online]. Available: {http://www.cstr.ed.ac.uk/projects/festival.html}

[12] V. Karaiskos, S. King, R. Clark, and C. Mayo, "The blizzard challenge 2008," in *in Proc. Blizzard Challenge workshop 2008*, 2008.

[13] S. King and V. Karaiskos, "The blizzard challenge 2011," in *in Proc. Blizzard Challenge workshop 2011*, 2011.

[14] "http://www.synsig.org/index.php/Main_Page," 2010.

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