DEVELOPMENT OF SYSTEM-BASED FEATURES FOR THE SPOOFED SPEECH DETECTION (SSD) TASK

Ms.Mandeep Kaur[1], Mr. R.K. Sharma[2]
[1,2]Guru Kashi University, Talwandi Sabo

ABSTRACT

In order to determine system characteristics, we investigate the basic concept that speech is analysed by the human ear in subbands (due to the signal processing abstraction of the cochlea, i.e., vibration of basilar membrane in a specific region for a specific tone). System-based features, such as the Sub band Auto encoder (SBAE) characteristic, which is an AE architecture modified to accommodate the human perception process, are employed in this article to recognise natural and faked speech. The Sub band Auto encoder (SBAE) function detects real and faked speech. It is an AE design that has been adapted to accommodate the human visual mechanism.

KEYWORDS: Sub bands, Auto encoder, Spoofed, Speech, and System.

I. INTRODUCTION

As illustrated in Figure 1, the generic SSD system may be separated into four components: the database, feature extraction, classification, and decision making.

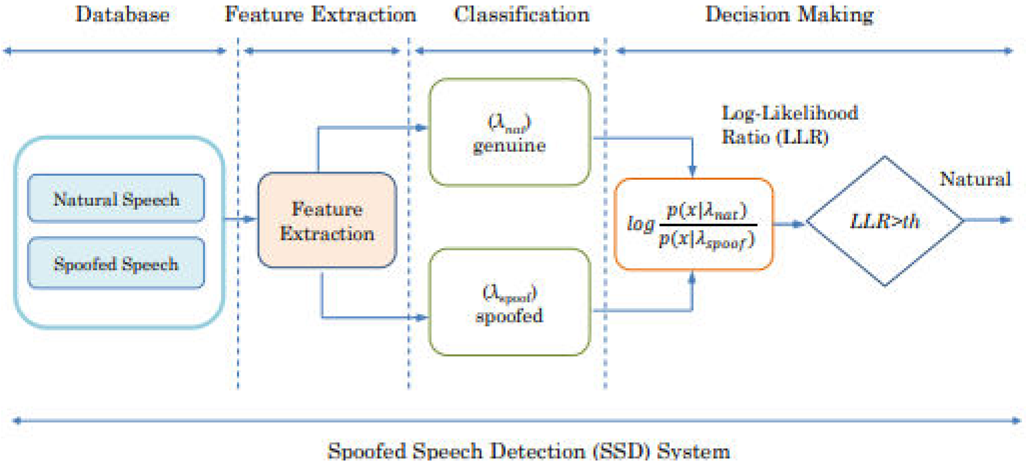

Database
Feature Extraction
Classification
Decision Making
Natural Speech
Spoofed Speech
Feature Extraction
(λ_{nat}) genuine
(λ_{spoof}) spoofed
Log-Likelihood Ratio (LLR)
$log \frac{p(x|\lambda_{nat})}{p(x|\lambda_{spoof})}$
LLR>th
Natural
Spoofed Speech Detection (SSD) System

Figure 3.5: General architecture of spoof detection system

A further point is that The human speech production system does not produce speech frame by frame (rather, it produces speech on a continuum that implicitly captures the naturalness of the speech production mechanism), whereas feature extraction in SS and VCS is typically done at the frame level. As a result, dynamic fluctuations between frames are critical for SSD task performance. The detection of SS speech is a vital need because any random text may be generated for any speaker, and in the event of VCS spoof, any speaker can be attacked (i.e., male-to-female and vice versa attacks are feasible).

II. TASK DESCRIPTION

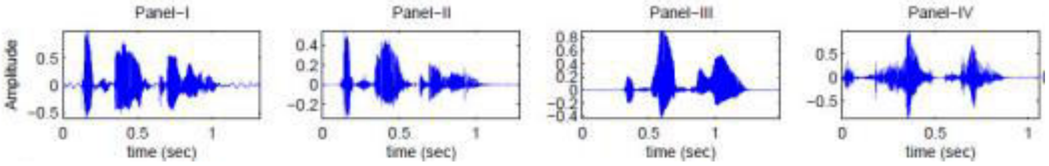
As a comparative job for spoofing speech detection, the ASVspoof 2015 challenge is intended to serve as a benchmark for participants. Three data sets are provided, comprising training, development, and assessment sets, as well as the statistics associated with each of these sets. There are a total of ten spoofing kinds, numbered S1 through S10. The first five types (S1-S5) are present in all data sets and are referred to as "known types" since they are commonly employed in system development. The final five types (S6-S10) are "unknown kinds," which means that they only exist in the evaluation set. S1) frame selection based VC; S2) VC that alters the first Mel cepstral coefficient; S3) HMM-based TTS customised to target speaker using 20 sentences; S4) same as S3 but using 40 adaption sentences per speaker; and S5) GMM-based VC taking global variance into account. In light of the fact that this is a detection issue, the official system performance metric for ASVspoof 2015 is the equal error rate (EER). The EER is the rate that occurs when the false alarm rate equals the miss rate. In order to acquire a more complete picture of processor speed, we additionally employ the detection error trade-off (DET) curve for system evaluation on development data for which we know the ground truth of whether a sentence is genuine or faking, in addition to the detection error trade-off curve. Speech synthesis and voice conversion spoofing attacks are both possible included in the database, as they are the most accessible and extremely successful spoofing tactics available today.

III. RESEARCH METHODOLOGY

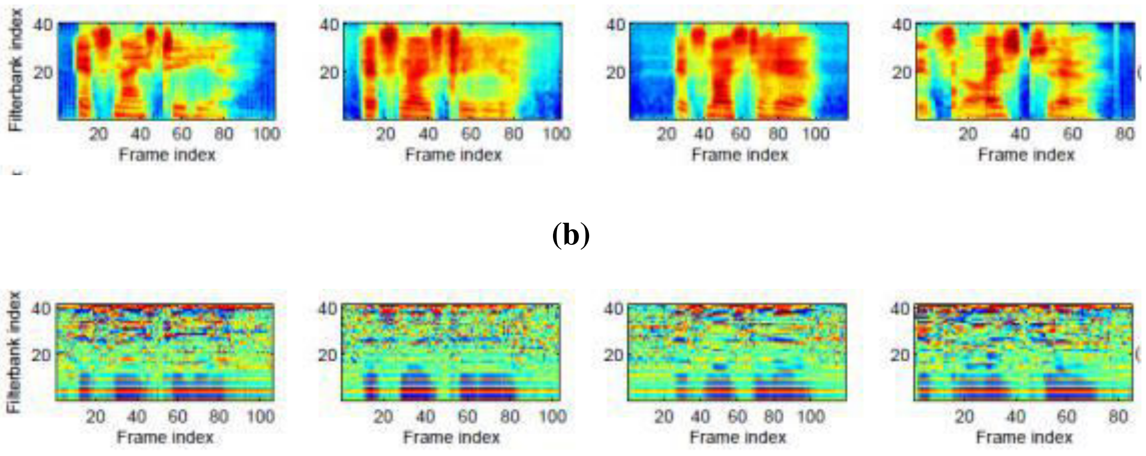
For the purpose of SBAE feature extraction, the voice signals were separated into frames with a 25-millisecond duration and 50-percent overlaps. SBAE was utilised to extract features from the STRAIGHT spectrum using the SBAE algorithm. For training purposes, all input and output data were normalised between 0 and 1. The SBAE was used to extract features from validation and evaluation datasets after being trained on training data., which were then fed back into the SBAE. In this case, 40 units in the sub band layer result in 40 sub band characteristics. For the discrimination test, SBAE traits related to higher sub bands are taken into account. The average value of two successive sub band characteristics was used to reduce the dimensionality of the data even further, and 24 sub bands were reduced to 12 sub bands in the process.

IV. DATA ANALYSIS

For the SSD job, it is necessary to observe the SBAE representation for both Figure 1 depicts natural and various varieties of faked speech 2 to find the impact of SBAE characteristics.



(a)



(b)

(c)

Figure 2: (a) Speech signal waveform, (b) Mel filterbank energies and (c) SBAE features energies for Panel I: natural speech, Panel II: vocoder-based VCS, Panel III: vocoder-based SS and Panel IV: USS-based MARY TTS.

It can be seen that both the MFCC and SBAE features change for natural and spoofs of varied types and sophistication. As a result, both of these features can be utilised to the problem of spoof detection. Furthermore, because both feature sets are invertible, the speech spectrum may be rebuilt by merging both sets of features (while it may not be strictly necessary for classification problem). The average Log Spectral Distortion (LSD) between the original spectrum P() and the reconstructed spectrum P() was calculated to assess both features' ability to reconstruct their original spectra. derived:

$$LSD = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{P(\omega)}{\hat{P}(\omega)}\right]^{2} d\omega} \ .$$

The LSD for 50 natural utterances of the ASVspoof 2015 database was 5.01 dB in the case of SBAE features, and 9.04 dB in the case of Mel filterbank energies. For natural speech and for spoofing assaults, the suggested features show little fluctuation in low-frequency areas, as shown in Figure 2. Due to nonlinear processing, it has been suggested that the existing characteristics are more sensitive to changes in the spectrum. If you look at two successive frames in Figure 1, you may see this impact. There is a greater variation in proposed features between consecutive frames compared to Mel filterbank energies (in timedomain). Thus, SBAE characteristics can collect more dynamic speech spectrum data.

As shown in Figure 3 (a) and Figure 3 (c), the most essential aspect of the SBAE feature for spoof identification is the difference in SBAE characteristics and Mel filterbank energies between real utterances and utterances synthesised with USS (S10 system in ASV spoof challenge database). There are various ways to create a spoof, but none are as difficult to detect as USS-based ones. The USS generates an output speech signal that corresponds to the text input in order to generate an output speech signal that corresponds to the text input. system concatenates distinct units of genuine speech. For this reason, it can be difficult to distinguish the difference between USS-generated speech and natural speech, because USS systems employ parts of natural speech. Modern features such as MFCC, which are effective against other types of attacks like VCS and HMM-based SS, are therefore ineffective against synthetic speech generated by USS systems. (a) And (b) show this impact, respectively (c). Natural speech and USS-based speech have nearly same variance in higher-order Mel filterbank energies. Other types of speech, such VCS (Figure 3b) and SS, show slightly different variances in the Mel filterbank energies. When it comes to spoofed speech of all kinds, higher-order SBAE traits exhibit more variation. Speech synthesised using USS can clearly be distinguished from natural speech. As a result, SBAE features may be preferable to MFCCs when it comes to USS speech recognition. Furthermore, because of the extremely high and low variance of the SBAE features for various types of spoof, it is anticipated that the performance will be enhanced than using only static information by integrating their dynamic changes as countermeasure.

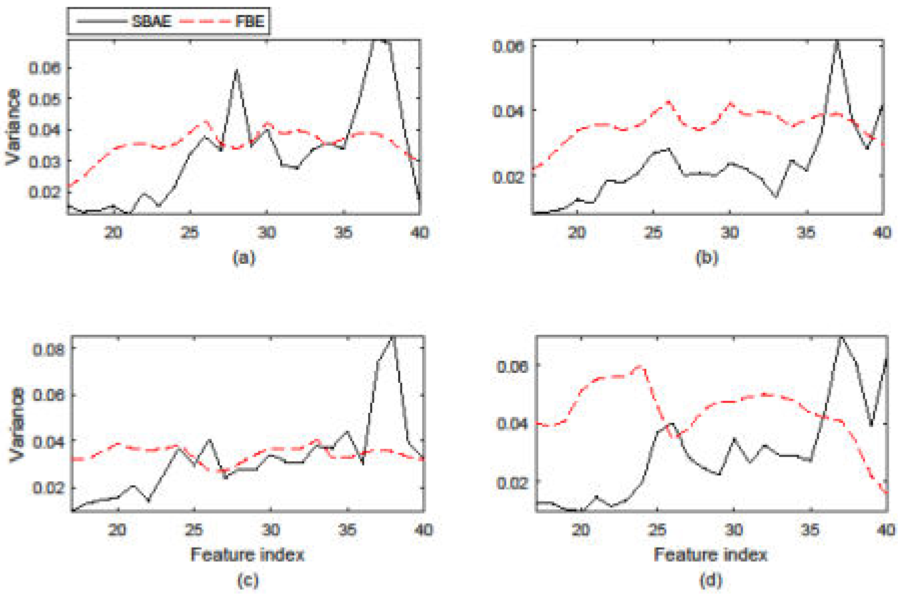


Figure 3: Variance of higher-order Mel filterbank energies (FBEs) and SBAE characteristics for (a) natural speech, (b) vocoder-related VCS, (c) USS-based MARY TTS speech and (d) vocoderrelated SS.

V. EXPERIMENTAL RESULTS

Table 1 shows the outcomes of the MFCC and SBAE development sets. According to the results, the static feature vector of the SBAE features gave an EER of 5.38 percent, which is greater than the EERs of the MFCC, CFCC, CFCCIF, and CFCCIFS feature sets. study. SBAE's EER is nearly identical to those of MFCC and CFCC when the characteristics are combined with the static features, i.e. 2.50 percent, 2.25 percent, and 2.59 percent, respectively. In addition to MFCC, CFCC and CFCCIF features alone, the EER of SBAE is reduced to 1.60 percent with the application of features. SBAE feature EER is reduced by using dynamic features, which captures more spectral variance than descriptors.

Table 1: Score-level fusion of SBAE with the MFCC and CFCC feature sets utilising the D1, D2, and D3 feature vectors at various fusion factors f on the development set yields EER (in percent) in terms of percentages.

Feature Set1	**Fusion Factor (α_f)**											**Feature Set2**
	0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	
SBAE: D1	5.50	4.50	3.80	3.30	3.00	2.72	2.59	2.80	2.95	3.13	3.26	**MFCC:D1**
SBAE: D2	2.50	2.00	1.63	1.44	1.50	1.40	1.59	1.64	1.80	2.11	2.25	**MFCC:D2**
SBAE: D3	1.60	1.20	1.00	**0.80**	**0.80**	0.77	1.00	0.9	1.26	1.46	1.60	**MFCC:D3**
SBAE: D1	5.50	4.65	4.00	3.60	3.25	3.20	3.19	3.59	4.13	4.35	4.55	**CFCC:D1**
SBAE: D2	2.50	2.00	1.70	1.39	1.50	1.55	1.63	1.82	2.15	2.49	2.60	**CFCC:D2**
SBAE: D3	1.60	1.15	1.00	**0.85**	0.95	1.00	0.97	1.12	1.19	1.50	1.54	**CFCC:D3**
SBAE: D1	5.50	4.60	3.80	2.99	2.40	2.12	1.97	2.21	2.11	2.11	2.29	**CFCCIF: D1**
SBAE: D2	2.50	1.90	1.45	1.23	1.20	0.9	1.00	1.13	1.23	1.39	1.40	**CFCCIF: D2**
SBAE: D3	1.60	1.35	1.12	0.90	**0.85**	0.91	0.94	1.13	1.23	1.42	1.52	**CFCCIF: D3**

SBAE: D1	5.50	4.49	3.65	2.72	2.11	1.82	1.69	1.76	1.81	1.85	1.89	**CFCCIFS: D1**
SBAE: D2	2.50	1.88	1.40	0.9	0.92	0.89	0.86	1.00	1.21	1.23	1.06	**CFCCIFS: D2**
SBAE: D3	1.60	1.10	1.00	0.80	**0.70**	0.80	0.74	0.99	1.13	1.11	1.23	**CFCCIFS: D3**

A table of findings is presented in Table 1 for the score-level fusion of SBAE and system-based features. Conventional 36-D AE features, on the other hand, yielded an EER of 8.1 percent. As a result, AE traits are not considered for further study in this work. In comparison to the EER achieved utilising only MFCC and SBAE characteristics, it is found that when f = 0.3 or 0.4, the fusion factor, the EER was 0.80 percent. SBAE features, on the other hand, were able to capture more information that MFCC features could not. The CFCC and CFCCIF features also yielded similar findings. The lowest EER of 0.63 percent is achieved when CFCCIFS characteristics are fused with SBAE features on the score level. Nonetheless, the CFCCIFS spoof detection features are to blame for this drop.

5.1 Outcomes on the Blizzard Challenge 2014 Database

Table 2 shows the performance of SBAE characteristics for based on data from the Blizzard Challenge 2014 database. Among the HMM-based systems tested for the Gujarati language, the HMM-based systems (D and H) were classified as having less than 10% EER, but the other HMM-based systems (C, E, and F) were classified as having 10-50% ERR. The utilisation of dynamic information improves performance. of the USS-based system G was greatly enhanced compared to before. Comparing the EER of the MFCC features to the SBAE features, the EER of the MFCC features is significantly lower. For Hindi, practically all HMM-based systems (with the exception of E) produced lower percent EER, while the As the amount of dynamic information increased, the performance of USS-based system G declined. Surprisingly, the HMM-DNN-based system for Gujarati achieved an EER of 47 percent, whereas the Hindi system only produced an EER of 5 percentpercent. When evaluating SBAE features on the ASV spoof challenge database, it was discovered that the reduction in EER was not consistent with the dynamic features in the case of SBAE features. SBAE features, on the other hand, achieved a lower percent EER for HMM-based speech than MFCC or other cochlear-based features, which was not the case with MFCC or other cochlear-based features. These qualities were unable to detect an extremely unknown attacking situation, and they are also somewhat dependent on the language used to determine the attack scenario. The experimental findings, on the other hand, did not support this conclusion.

Table 2: Training using ASV spoof data and examine with Blizzard Challenge databases resulted in an EER (in percent) for the SBAE feature set utilising D1, D2, and D3 feature vectors on the SBAE feature set.

Blizzard2012					**Blizzard2014**					**Blizzard2014**				
English		**SBAE**				**Gujarati**	**SBAE**				**Hindi**	**SBAE**		
		D1	**D2**	**D3**			**D1**	**D2**	**D3**			**D1**	**D2**	**D3**
USS	**B**	45	45	42	**C**	**HMM**	60	34	47	**B***	**HMM**	20	3	2
Hybrid	**C**	40	35	40	**D**	**HMM**	10	50	2	**C**	**HMM**	6	0	2
Hybrid	**D***	50	72	75	**E**	**HMM**	60	87	90	**D**	**Hybrid**	28	5	5
HMM	**E***	10	12	25	**F**	**HMM-DNN**	80	50	70	**E**	**HMM**	70	4	32
USS	**F**	72	65	64	**G**	**USS**	85	18	5	**F**	**HMM-DNN**	18	6	7
USS	**G**	25	80	80	**H**	**HMM**	47	10	20	**G**	**USS**	40	45	54
HMM	**H**	15	45	55						**H***	**HMM**	34	20	15
USS	**I**	38	60	42						**K**	**HMM**	3	0	0
Diphone	**J***	59	32	40										
HMM	**K***	44	60	62										

VI. CONCLUSION

When it came to the problem of spoof detection, we looked into a suggested SBAE function. The feature sets produced good results when tested against the ASV spoof 2015 database, particularly when tested against unknown attacks. Dynamic characteristics had a substantial impact to the overall performance, but the type of spoof and channel fluctuations played a role in addition These inferences were derived from system-based attributes, with no element directly including excitation source-based features. The absence of source information is problematic. another factor contributing to the unnaturalness of the spoof speech.

REFERENCES

1. H. A. Patil, "Speaker Recognition in Indian Languages: A Feature Based Approach," Ph.D. Thesis, Dept. of Elec. Engg., IIT Kharagpur, 2005.

2. H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Comm., vol. 27, no. 3-4, pp. 187-207, Apr. 1999.

3. G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504-507, 2006.

4. N. Jaitly and G. E. Hinton, "A new way to learn acoustic events," Advances in Neural Information Processing Systems, vol. 24, pp. 1-9, 2011.

5. M. H. Soni, T. B. Patel, and H. A. Patil, "Novel sub bandauto encoder features for detection of spoofed speech," in Int. Speech Comm. Assoc. (INTERSPEECH), San Francisco, USA, 2016, pp. 1820-1824.

6. N. Jaitly and G. E. Hinton, "Using an auto encoder with deformable templates to discover features for automated speech recognition," in Int. Speech Comm. Assoc. (INTERSPEECH), Lyon, France, 2013, pp. 1737-1740.

7. A Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The det curve in assessment of detection task performance,” DTIC Document, Tech. Rep., 1997.

8. Z. Wu, et al., "Spoofing and counter measures for speaker verification: A survey," Speech Comm., vol. 66, pp. 130-153, 2015.

9. A Neustein and H. A. Patil, Forensic Speaker Recognition: Law Enforcement and CounterTerrorism. Springer, Oct. 2011.

10. D. Yambay, J. S. Doyle, K. W. Bowyer, A. Czajka, and S. Schuckers, "LivDet-Iris 2013 - Iris liveness detection competition 2013," in IEEE Int. Joint Conf. on Biometrics (IJCB), Florida, USA, 2014, pp. 1-8.