Spoofing Countermeasure Based on Analysis of Linear Prediction Error

Artur Janicki

Institute of Telecommunications, Warsaw University of Technology, Warsaw, Poland

A.Janicki@tele.pw.edu.pl

Abstract

In this paper a novel speaker verification spoofing countermeasure based on analysis of linear prediction error is presented. The method analyses the energy of the prediction error, prediction gain and temporal parameters related to the prediction error signal. The idea of the proposed algorithm and its implementation is described in detail. Various binary classifiers were researched to separate human and spoof classes. When tested on the corpora provided for the ASVspoof 2015 Challenge, the proposed countermeasure yielded much better results than the baseline spoofing detector based on local binary patterns (LBP). It is hoped that the proposed method can help in developing a generalised countermeasure able to detect spoofing attacks based on different variants of speech synthesis, voice conversion, and, potentially, also other spoofing algorithms.

Index Terms: speaker verification, spoofing, linear prediction, local binary patterns, binary classification

1. Introduction

Automatic speaker verification systems (ASVs), which use the human voice to authorise user access, are becoming more and more widely used. Since other speech processing algorithms, such as voice conversion or speech synthesis, are becoming easily available and improving their quality, they have started to pose a major threat to ASV systems.

Quite recently, researchers have started to investigate how much ASV systems are prone to spoofing. Various researchers have worked on assessing the threat caused by imitators [1],[2], synthetic speech [3],[4], converted speech [5],[6] or replay of previously acquired recordings [7], [8]. In parallel, much effort has been invested in elaborating various spoofing countermeasures, which were either dedicated to a given attack or claimed to be generally applicable. A thorough review of spoofing methods and their countermeasures can be found in [9].

The work described in this paper aims to contribute to the speech community's efforts to find efficient anti-spoofing methods for ASV systems. The experiments further described were conducted using the datasets that were made available by the Organisers of ASVspoof 2015 – the first ASV spoofing and countermeasures challenge [10], to be held during Interspeech 2015 in Dresden, Germany. This initiative aims to motivate the community to elaborate new, effective spoofing countermeasures. The datasets provided contain recordings of various access trials, partially annotated either as human voice or as a spoof trial, generated using one of 10 different speech synthesis or voice conversion algorithms. One corpus was not annotated and was supposed to contain also trials generated using previously unseen spoofing algorithms. Participants were requested to submit their scores, and they got back the evaluation of the efficiency of their countermeasures.

1.1. Aim of this work

In this work we present a novel countermeasure against spoofing using voice conversion, speech synthesis and, potentially, also other spoofing methods. The proposed method is based on analysis of prediction error. We were motivated by the fact that synthetic or converted voice is quite likely to be either very easily predicted, if generated with a simplified acoustic model, or very difficult to predict, if any artefacts in the signal are present. The results of experiments with the proposed countermeasure will be compared with the results of the detector based on local binary patterns (LBPs), which have turned out to be efficient in other studies.

In this paper we will first briefly present the state-of-the-art spoofing countermeasures, we will recall the principles of linear prediction theory, and then, in Section 3, we will present details of the proposed countermeasure. In Section 4 we will describe the experimental setup. Section 5 will present the results and discussion and, finally, Section 6 will conclude the paper.

2. Previous work

2.1. Spoofing countermeasures for ASV systems

Several countermeasures exist which exploit prior knowledge about the origin of the spoofing attack. For example, there are algorithms which try to detect artefacts which are likely to appear in speech synthesis, such as simplification of F0 contours [11]. In [12] the authors proposed an algorithm which is based on measuring the pair-wise distance (PWD) between spectral parameters (such as LPCs or MFCCs) in consecutive frames. The authors claimed that voice conversion causes a decrease in PWD values and, as a consequence, in PWD distributions. They compared speaker-dependent PWD distributions between genuine and converted speech, using speech data from the NIST'06 database and the NIST SRE protocol. The authors showed that the proposed countermeasure is able to lower the equal error rate (EER) from more than 30% to below 3%.

Countermeasures dedicated to detecting replay attacks often try to identify unexpected channel artefacts indicative of recording and replaying. Such algorithms were reported in [13], for which the EER for a baseline GMM-UBM system was shown to decrease from 40% to 10% with active countermeasures. Another replay countermeasure aimed at detecting far-field recordings, which are unlikely in natural access scenarios [14].

Only a few algorithms claim to be less dependent on prior knowledge of the attack. Such an approach was described in [15]. It was based on the local binary pattern (LBP) analysis of speech cepstrograms and was inspired by the original application to image texture analysis [16]. In this approach LBP analysis was applied to a mel-scaled cepstrogram with appended dynamic features. The authors claimed that mod-

ifications made through spoofing disturb the natural 'texture' of speech. Experimental results presented in [15] showed that the LBP-based textrogram analysis was effective in detecting spoofing trials generated using speech synthesis (EERs of below 1%), but it was less effective in detecting those originating from voice conversion (EER in the order of 7%).

Another generalised method is based on the fact that many speech synthesis and voice conversion algorithms disturb the natural phase of the speech signal. In [17] the authors challenged GMM-UBM and SVM-GMM speaker verification systems with genuine and synthesised speech originating from the WSJ corpus. They showed that by using relative phase shift (RPS) features it was possible to decrease EER from over 81% to less than 3%. Unfortunately, the method proposed was vocoder-dependent. Similarly, phase information was successfully used in detecting converted speech in [18].

2.2. Linear prediction theory

The linear prediction technique is a fairly old technique dating back to the 1940s [19]. It has been used not only in speech processing, but also in neuroscience and geology [20]. In speech processing it was originally used in speech coding, where a technique called linear prediction coding (LPC) was developed. Its idea consists in calculating the so-called prediction coefficients a_i so that in a frame (e.g., 20 ms long) of the input speech signal, each signal sample $x(n)$ can be efficiently predicted using the p preceding samples by the value $\hat{x}(n)$ calculated as follows:

$$\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i) \tag{1}$$

The difference between the original signal $x(n)$ and predicted signal $\hat{x}(n)$ is called the prediction error $e(n)$. The value G_p defined as

$$G_p = \frac{E_x}{E_e} \tag{2}$$

where E_x is the energy of signal $x(n)$ and E_e is the energy of prediction error $e(n)$, is called the prediction gain. The higher the gain is, the better the signal is predicted, so this means that the prediction coefficients $a_1..a_p$ were able to efficiently model the speech signal within a frame, which will allow for a better compression.

The LPC technique is widely used in speech coding, e.g., in GSM 06.10 [21] or in narrow-band and wide-band adaptive multi-rate coders (AMR) [22]. It can also be used to parametrise signal in speech or speaker recognition. Linear prediction can also be applied to vectors – in such a case, a vector of samples is predicted using another vector of samples from the signal's history. This method is called long-term prediction (LTP) and is often used on top of the LPC, i.e., LTC error is further processed by the LTP. This approach is encountered, e.g., in [21]. LTP works especially efficiently for voiced speech, where the signal is quasi-periodic. Prediction error and prediction gain are defined in the same way as for LPC.

3. Proposed countermeasure

The idea of the proposed ASV spoofing countermeasure is based on analysis of signal prediction error of the signal at the ASV input. One may expect that if a non-natural speech signal undergoes the prediction process, it may be either "too well" predicted (i.e., with a high prediction gain) or ineffectively predicted (i.e., with a prediction gain lower than usual).

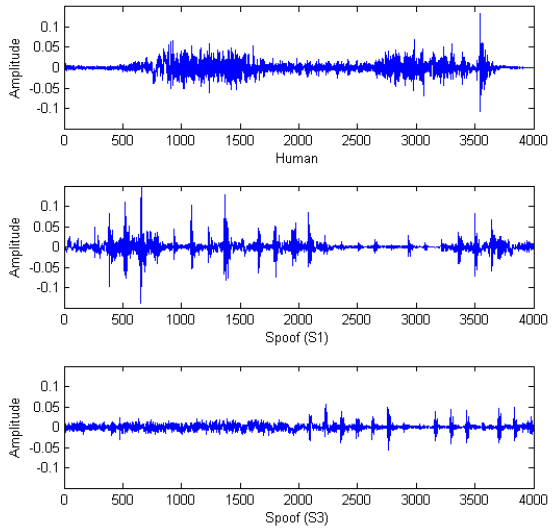

Figure 1: Residual signal ($x''(n)$) at the output of the LTP block for human speech (top), spoof signal S1 generated with voice conversion (middle) and spoof signal S3 generated with speech synthesis (bottom), all for voiced speech.

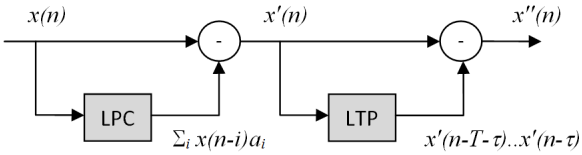

Figure 2: Schematic picture of speech processing in the proposed countermeasure.

Fig. 1 shows three residual (error) signals left from prediction processing of voiced speech. It can be observed that the prediction error of a spoof signal generated with voice conversion (middle) exhibits sudden bursts of errors, probably due to the non-smooth frame concatenation used in this case (algorithm S1, based on frame selection). These bursts are separated by a low energy noise-like prediction error signal, which may imply a much more efficient prediction than for natural human speech (top figure). The prediction error of the spoof signal generated with speech synthesis, shown in the bottom figure, is also much weaker and less dynamic than for natural speech. Therefore, in our approach we are going to measure various parameters of prediction error, hoping to capture the features which will help to differentiate human from spoof access trials.

The proposed speech processing process is shown in Fig. 2. It is similar to the speech coding process used, e.g., in GSM 06.10 coding [21]. The input signal $x(n)$ is first analysed using the LPC technique, where p prediction coefficients a_i are estimated. The predicted values are subtracted from the original samples, and the resulting LPC prediction error signal $x'(n)$ is processed further. The next block, LTP, operates on vectors of samples rather than on individual samples. When the best matching vector is found, it is subtracted from signal $x'(n)$, resulting in LTP prediction error signal $x''(n)$.

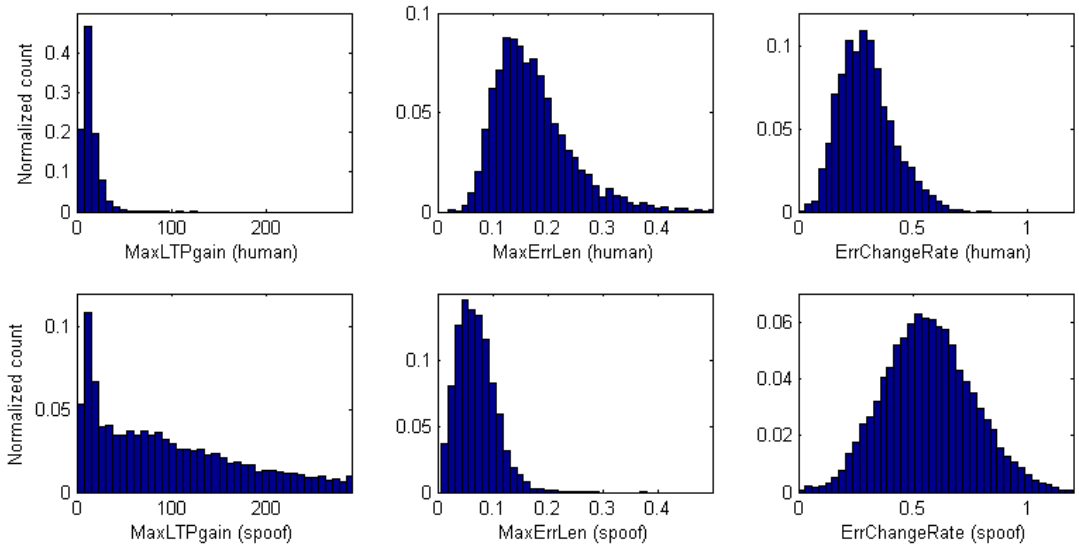

Figure 3: Histograms of selected features for human (top) and spoof (bottom) trials, for Training dataset (S2 excluded).

In our approach we analyse the energy of prediction errors resulting from both blocks and the ratios between them (i. e., LTP prediction gain). In addition, we analyse the temporal distribution of LTP errors, by measuring the length of segments with a prediction error above a certain level. In total, we propose to extract 10 parameters:

- *MeanLPCerr* – mean energy of the LPC error, i.e., mean energy of $x'(n)$;
- *MeanLTPerr* – mean energy of the LTP error, i.e., mean energy of $x''(n)$;
- *MaxLTPerr* – maximum energy of the LTP error;
- *MeanLTPgain* – mean LTP gain (i.e., mean ratio between energies of the LPC and LTP errors, mean G_p as defined in Eq. 2);
- *MaxLTPgain* – maximum of the G_p for LTP;
- *MeanErrLen* – mean length of segments with the LTP error above threshold θ;
- *MaxErrLen* – maximum length of segments with the LTP error above threshold θ;
- *MeanNoErrLen* – mean length of segments with the LTP error equal to or below threshold θ;
- *MaxNoErrLen* – maximum length of segments with the LTP error equal to or below threshold θ;
- *ErrChangeRate* – LTP threshold crossing rate (counted per 20ms frame).

Fig. 3 shows distributions of the three selected parameters: MaxLTPgain, MaxErrLen and ErrChangeRate, for both human and spoof trials. They show that for spoof trials MaxLTPgain reaches much higher values than for natural speech – probably due to the higher determinism of synthetic and converted voices. For the same reason, MaxErrLen values for human trials are higher, because the prediction errors for human speech not only have higher values, but also last longer, while for spoof trials they often take the form of short-term bursts caused by synthesis artefacts, separated by low energy error. In contrast, ErrChangeRate for human speech is usually lower, due to the lack of artefacts in natural speech generation.

4. Experimental set-up

The experiments were conducted on the corpora provided by the ASVspoof 2015 Challenge Organisers. They were divided into three parts: Training, Development and Evaluation, and consisted of 16,375, 53,372 and 193,404 recordings, respectively. The spoof trials were generated using 10 different spoofing algorithms (S1..S10), based either on speech synthesis or on voice conversion. Their spoofing efficiency ranged from 25.42% to 45.79% EER, with the exception of S2, which yielded a very low spoofing efficiency equal to 0.87% EER. The baseline EER value achieved with the PLDA system equalled 0.42% [10].

The proposed spoof detection system and the baseline LBP-based detector were trained using the Training database. Experiments with spoofing detection, including parameter tuning, were run using the Development corpus. The Evaluation corpus was tested to generate scores for the Challenge. No external data sources were used.

To increase granularity, both errors were calculated sample-by-sample and not frame-wise, which is the case in speech coding. Prediction order p and threshold value θ were set experimentally to 2 and 0.011, respectively. The analysis was narrowed to voiced regions of speech only, where linear prediction reaches the highest gain. Voicing detection was realised using the SWIPE pitch detector [23]. The proposed spoofing countermeasure will hereinafter be referred to as LPA (linear prediction analysis).

The baseline LBP-based countermeasure was set up according to the description in [15]. Each signal was analysed forming a feature matrix consisting of 16 cepstral coefficients plus energy, their deltas and delta-delta coefficients, which was further analysed using 58 possible uniform LBP patterns. As a result, 2842 features were generated for every recording.

We used a range of binary classifiers, trying to achieve the

largest area under the receiver operating characteristic (ROC) curve. Based on these results, three classifiers were selected: Logistic [24], Bayesian Networks [25] and AdaBoosts [26] classifier. Feature extraction was carried out in Matlab, using Voicebox[1] as a speech processing library. Experiments with classification were run using the WEKA toolkit [27].

5. Results

Spoofing detection was evaluated according to the challenge description, i.e., by measuring the EER values. Since the spoofing efficiency caused by method S2 was very low, some of the measurements were done on a subset without S2 trials, and therefore the EER results were denoted either as EER_{all} or EER_{noS2}, respectively. The EER values and DET plots were obtained using the Bosaris toolkit[2].

Table 1: *EER results (in percentages) of the spoof detection, for Development and Evaluation datasets, for various classifiers, for LBP and LPA countermeasure methods.*

Method	Metrics	Logistic	BayesNet	AdaBoost
		Development		
LBP	EER_{noS2}	11.529	30.168	12.496
	EER_{All}	14.791	30.618	15.795
LPA	EER_{noS2}	2.986	5.956	3.268
	EER_{All}	8.905	11.065	9.386
		Evaluation		
LPA	EER_{All}	11.616	-	-

The results presented in Table 1 show that the EER values achieved for the Development set with the proposed LPA countermeasure proved to be very efficient. The Logistic classifier yielded the best results, with less than 9% EER for the whole Development set and less than 3% EER for the same set without S2. The LBP-based detector achieved significantly worse results: more than 14% EER and more than 11% EER, respectively. The AdaBoost classifier returned slightly worse results than the Logistic classifier. Bayesian Networks yielded 11% and 6% EER, respectively, for the LPA detector, and over 30% EER for the LBP-based detector.

Since the Logistic classifier performed the best, it was used to generate scores for the Evaluation dataset for the challenge submission. Having evaluated the submission, the Organisers returned the EER of 11.6% for the whole Evaluation corpus – 6.1% for the known attacks and 17.1% for the unseen ones.

Fig. 4 presents the DET curves for the best classifier (Logistic), for the proposed method and the baseline LBP-based detector. The plot confirms that the proposed LPA detector performs better than LBP. The curves of both detectors for the whole set seem to converge for low false alarm values. The lines are approximately straight. It is noticeable that the distance between the lines for the whole set and for the set without S2 for both detectors is long. The DET plot indicates that for the Development corpus without S2 the proposed countermeasure was not able to decrease the miss probability below 0.16 %.

6. Conclusions

In this paper we presented a novel spoofing countermeasure which can be used to protect speaker verification systems from

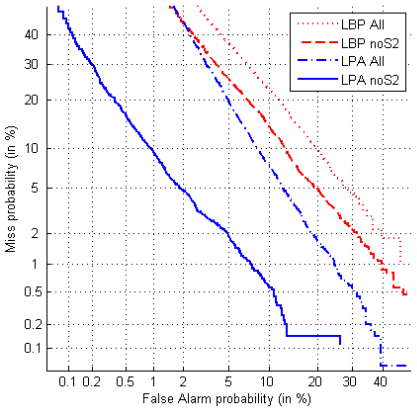

Figure 4: DET plots for the baseline (LBP-based) and the proposed (LPA) countermeasure tested on the Development set.

unauthorised access using speech synthesis, voice conversion, and, potentially, also other attacks. It proved to be very efficient when tested on the Development set – it yielded less than 3% EER on a subset without S2 (which was not effective in spoofing anyway).

The method is based on analysis of prediction error, resulting from cascaded LPC and LTP blocks. As the LP-based vocoder is often a part of speech synthesis or voice conversion systems, by using linear prediction analysis in a way we perform a reverse operation, to verify if vocoding really took place. Prediction gain values are analysed, as well as their temporal features, such as mean length of segments with low prediction error. The proposed method performed significantly better than a baseline LBP-based detector. It is likely that the LBP detector requires longer speech data (in [15] it was tested on 5 min. recordings), while the recordings tested in the current study were no longer than several seconds.

The results achieved with the LPA detector on the Evaluation corpus for the known spoofing attacks (five algorithms) were even better than for the Development corpus (6.1% vs. 8.9%). The EER for the whole Development set was worse – over 11%. This may imply that the previously unseen algorithms differ significantly from the algorithms used to generate the Training and Development datasets, so the proposed algorithm requires further parameter tuning. These results are anyway considered as promising, knowing that the EER under spoofing without any countermeasures used equalled 29.3%.

Further analysis of the results will be possible when the ASVspoof 2015 Organisers provide the keys to the Evaluation set. The authors hope that the proposed approach, based on analysis of prediction error, will help in the future to elaborate a generalised countermeasure able to precisely detect a wide range of spoofing attacks against ASV systems.

7. Acknowledgements

The calculations described in the article were made in the Interdisciplinary Centre for Mathematical and Computational Modelling (ICM) of the University of Warsaw (computational grant No. G46-2).

[1] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
[2] https://sites.google.com/site/bosaristoolkit/

8. References

[1] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on*. IEEE, 2004, pp. 145–148.

[2] M. Farrús, M. Wagner, J. Anguita, and J. Hernando, "How vulnerable are prosodic features to professional imitators?" in *Proc. IEEE ODYSSEY - The Speaker and Language Recognition Workshop*, 2008.

[3] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech," in *Proc. EUROSPEECH*, 1999.

[4] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA workshop*, 2010, pp. 131–134.

[5] T. Kinnunen, Z. Wu, K. A. Lee, F. Sedlak, E. S. 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 and Signal Process. (ICASSP)*, 2012, pp. 4401–4404.

[6] Z. Wu, A. Larcher, K.-A. Lee, E. Chng, T. Kinnunen, and H. Li, "Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints." in *INTERSPEECH*, F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier, Eds. ISCA, 2013, pp. 950–954.

[7] J. Lindberg and M. Blomberg, "Vulnerability in speaker verification - a study of technical impostor techniques," in *European Conference on Speech Communication and Technology*, 1999, pp. 1211–1214.

[8] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, Sept 2014, pp. 1–6.

[9] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communications*, vol. 66, pp. 130–153, 10 2014.

[10] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilc, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *in Proc. Interspeech 2015*, Dresden, Germany, 2015.

[11] A. Ogihara and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 88, no. 1, pp. 280–286, 2005.

[12] F. Alegre, A. Amehraye, and N. Evans, "Spoofing countermeasures to protect automatic speaker verification from voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2013.

[13] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, vol. 4, July 2011, pp. 1708–1713.

[14] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *Security Technology (ICCST), 2011 IEEE International Carnahan Conference on*, Oct 2011, pp. 1–8.

[15] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," Lyon, France, 2013.

[16] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[17] 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 and Signal Process. (ICASSP)*, 2011, pp. 4844–4847.

[18] Z. Wu, E. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. 13th Interspeech*, 2012.

[19] P. P. Vaidyanathan, *The Theory of Linear Prediction*, ser. Synthesis Lectures on Signal Processing. Morgan & Claypool Publishers, 2007.

[20] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[21] *Digital cellular telecommunications system (Phase 2+) (GSM); Full rate speech; Transcoding GSM 06.10*, ETSI Std., Rev. version 8.1.1, 1999.

[22] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jrvinen, "The adaptive multirate wideband speech codec (amr-wb)." *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, 2002.

[23] A. Camacho, "Swipe: A sawtooth waveform inspired pitch estimator for speech and music," Ph.D. dissertation, Gainesville, FL, USA, 2007.

[24] J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2 2010.

[25] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[26] Y. Freund and R. E. Schapire, "A short introduction to boosting," in *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 1999, pp. 1401–1406.

[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.