INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

A Hierarchical Predictor of Synthetic Speech Naturalness Using Neural Networks

Takenori Yoshimura[1], *Gustav Eje Henter*[2], *Oliver Watts*[2], *Mirjam Wester*[2], *Junichi Yamagishi*[2,3], *and Keiichi Tokuda*[1]

[1]Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Japan
[2]The Centre for Speech Technology Research, The University of Edinburgh, UK
[3]National Institute of Informatics, Tokyo, Japan

{takenori,tokuda}@sp.nitech.ac.jp, {ghenter,owatts,mwester,jyamagis}@inf.ed.ac.uk

Abstract

A problem when developing and tuning speech synthesis systems is that there is no well-established method of automatically rating the quality of the synthetic speech. This research attempts to obtain a new automated measure which is trained on the result of large-scale subjective evaluations employing many human listeners, *i.e.*, the Blizzard Challenge. To exploit the data, we experiment with linear regression, feed-forward and convolutional neural network models, and combinations of them to regress from synthetic speech to the perceptual scores obtained from listeners. The biggest improvements were seen when combining stimulus- and system-level predictions.

Index Terms: speech synthesis, naturalness, neural network, Blizzard Challenge

1. Introduction

There is still no well-established objective measure to automatically quantify the naturalness of the synthetic speech generated by text-to-speech (TTS) systems. This contrasts with other areas of speech technology research, *e.g.*, automatic speech recognition (ASR), where word error rate is the standard performance measure, and speaker recognition, where equal error rate is commonly used to judge overall system accuracy. Automatically computed measures of the naturalness of synthetic speech have, of course, been proposed: mel-cepstral distance (MCD) [1] and root mean squared error (RMSE) of fundamental frequency (F_0) are widely used in the speech synthesis community. However, such measures often do not correlate well with human perception. This means that engineers must conduct expensive and time-consuming subjective evaluations, where dozens of human participants provide numeric ratings of synthetic speech. Intensive tuning of TTS system parameters is therefore difficult. Furthermore, standard techniques for system training offer no guarantees that synthetic speech is natural in terms of human perception, since TTS systems are usually optimized using maximum likelihood or minimum error criteria, rather than a criterion which is perceptually defined.

To overcome these problems, researchers have introduced measures proposed in telecommunications research [2, 3, 4, 5] and proposed new objective measures [6, 7] as alternatives to conventional TTS performance measures. Despite these efforts, most researchers still use the conventional measures to evaluate synthetic speech, resulting in poor correlation with human perception. Rather than manually crafting perceptual measures, an alternative approach is to use machine learning to generalize from a database of listener judgements. The Blizzard Challenge (BC) [8] data is suitable for this purpose. In the Challenge, participants must build a synthetic voice from the released speech database and synthesize a given set of test sentences. The sentences from each synthesizer are then evaluated through large-scale listening tests. Data from six years of the challenge [9, 10, 11, 12, 13, 14] – including the synthetic speech output of many synthesizers, natural speech, and listener responses – have been made publicly available by the organizers. Some researchers have already tried to predict speech-quality using acoustic features extracted from BC data [3, 15, 16, 17]. We also follow this approach, but rather than incorporating a large number of hand-engineered acoustic features into predictors, we instead propose extracting relevant features automatically as part of model training. This is expected to overcome some of the limitations of conventional measures:

1) Frame-wise nature: Global patterns such as the F_0 contour over the course of a syllable are ignored in spite of their importance to naturalness.

2) Local vs. global degradation: While conventional measures can capture global degradation by computing averages over all frames, it is difficult to handle local problems, *e.g.*, spectral discontinuities at phone boundaries and a sudden, inappropriate F_0 excursion on a certain word, which often dominate subjective naturalness perception [18].

Our predictors automatically learn features at various levels using convolutional neural networks (CNNs) to effectively overcome the above-mentioned problems. CNNs have the capacity to capture both local and global degradations through their convolutional-pooling layers that operate at different levels of detail. In addition, NNs are used for regression instead of the linear regression techniques typically used in previous work [15]. Furthermore, we combine several speech-quality predictors by integrating stimulus- and system-level predictions. Although the two predictions are performed independently in previous work [15], we stack them to make use of system-level knowledge for stimulus-level prediction, and to train both predictors simultaneously.

NNs have been used previously for automatic evaluation of low bit-rate audio coding schemes [19] and CNNs have been found to be effective for the automatic assessment of the perceived quality of digital videos [20]. As far as we know, however, this is the first attempt to use NNs for predicting the naturalness of synthetic speech.

The remainder of the paper is organized as follows: In Sec-

http://dx.doi.org/10.21437/Interspeech.2016-847

tion 2, the dataset used for our experiments is described, and the predictability of its ratings is investigated. Sections 3 and 4 report on experiments using NNs and CNNs, respectively. Conclusions and future work are presented in Section 5.

2. Data used and its predictability

2.1. Data

The BC annually conducts large-scale perceptual experiments where several hundred stimuli, generated by multiple synthesizers, are rated from 1 (bad) to 5 (excellent) by more than 200 human listeners. For all experiments reported in this paper, we used the BC data from six years [9, 10, 11, 12, 13, 14]. One English task with all the domains (except reportorial and paragraph domains) for each year was used; details are given in Table 1. This paper takes mean opinion score (MOS) for naturalness as the subjective measure of interest, although the methods devised are expected to be applicable to other subjective scores.

2.2. Inherent predictability of the data

An important concern when working with speech naturalness prediction is whether subjective scores are inherently predictable or not. To verify that listener ratings of synthetic speech are indeed predictable, a bootstrap method [21, 22] was applied to the database: By random sampling with replacement from the experimental data and measuring how the resampled MOSs vary, we can estimate how much the subjective scores would change if we were to call in an equally-large set of similar but independent listeners. The following procedure was used:

1. Compute the set of observed MOSs, $\boldsymbol{M}$, from the original subjective evaluation data.
2. $b \leftarrow 1$, where b is a replication index.
3. Randomly draw N listeners with replacement from the N original listeners in the subjective evaluation.
4. Using the scores of the drawn set of listeners, compute the set of MOSs in bootstrap replication b: $\boldsymbol{M}^{(b)} = [M_1^{(b)}, \ldots, M_S^{(b)}]$, where S is the number of stimuli and $M_s^{(b)}$ is the MOS of a stimulus s in replication b.
5. Compute the correlation coefficient $\rho^{(b)}$ and other similarity measures between $\boldsymbol{M}$ and $\boldsymbol{M}^{(b)}$
6. $b \leftarrow b + 1$.
7. If $b \leq B$ then go to step 3.
8. Compute basic statistics of the set $\{\rho^{(b)}\}$ and other similarity measures across all b.

Here, $B = 1000$ and the MOS of natural speech was excluded. BC listening tests follow a balanced design in which listeners are assigned to groups, and each listener group hears the same exact set of audio stimuli. The number of listeners from each group in each bootstrap replication b was the same as in the original data. For each replication, four similarity measures were computed with reference to the original data: mean absolute error (MAE), RMSE, Pearson's correlation coefficient (ρ), and Spearman's rank correlation coefficient (ρ_s). For each measure, four basic statistics – mean, standard deviation (SD), minimum, and maximum – were computed over the B replications and averaged across all years.

Table 2 shows the results of the bootstrap experiment. The MAE and the RMSE were around 0.2 and the correlations were adequately high ($\rho, \rho_s > 0.9$). This indicates good agreement among listeners. Thus, MOS for naturalness is predictable, at least in the BC data used in this paper.

We also used the data replications to investigate the sensitivity of speech naturalness prediction to variation in the response data used for training. To this end, we replicated the training of a predictor from previous work [15], which uses principal component regression to regress from a set of features used in the P.563 measure [2] to a MOS for naturalness. For each of the B bootstrap replications, two predictors were trained, on the bootstrap replication and on the original subjective test data. The differences and the correlation coefficients between MOSs predicted by the two systems were calculated in the same manner as in the previous experiment. The differences and the correlation coefficients were nearly 0.0 and 1.0, respectively. Predictors derived from subjective tests seem to be insensitive to variation among listeners in the training data.

3. Speech naturalness prediction using NNs

The dataset available from past BCs is much larger than that obtained from a typical subjective experimental setup, where, *e.g.*, 50 stimuli might be evaluated by as few as 10 listeners. We expect NNs to be able to outperform conventional linear regression methods where such large-scale data is available.

3.1. Prediction framework

There are two types of scores which have been predicted for synthetic speech in previous work [15]:

- System-level score: Overall performance of a synthesizer, averaged over all listeners and stimuli for a system.
- Stimulus-level score: Overall quality of an utterance or several utterances, averaged over all listeners from whom responses are available for these utterances.

We devised a framework for speech naturalness prediction to improve the prediction of both these measures. Figure 1 illustrates a standard prediction system, such as the one used in [15]. In this framework, a stimulus-level feature is extracted from speech, and a stimulus-level score obtained by feeding these features into a predictor. A system-level score is then obtained

Table 1: *Summary of the BC data used.*

	BC 2008	BC 2009	BC 2010
Task	full	EH1	EH1
Domain	news novel	news conversational	news novel
# Systems	20	17	17
# Stimuli	840	663	612
	BC 2011	BC 2012	BC 2013
Task	EH1	EH2.1	EH1
Domain	news novel	news novel	news novel
# Systems	12	10	9
# Stimuli	312	420	477

Table 2: *Differences and correlation coefficients between MOSs derived from bootstrap replications and the original subjective evaluation result.*

	Mean	SD	Min	Max
MAE	0.18	0.01	0.14	0.23
RMSE	0.24	0.02	0.19	0.31
ρ	0.96	0.01	0.93	0.98
ρ_s	0.96	0.01	0.93	0.97

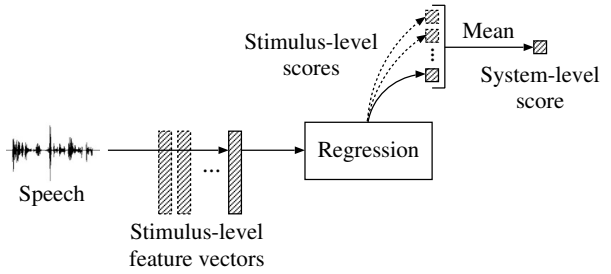


Figure 1: *Standard prediction system: a predicted system-level score is computed by averaging predicted stimulus-level scores.*

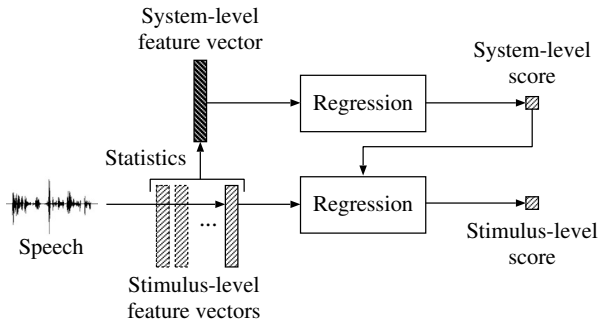


Figure 2: *Hierarchical prediction system: the predicted system-level score is used as a feature for stimulus-level prediction.*

by averaging all the predicted stimulus-level scores. However, previous papers [15, 17] have reported large prediction errors for stimulus-level predictions, but shown that system-level scores can be reasonably predicted with low RMSE and high correlation with listeners' ratings. Drawing inspiration from this, we here propose an alternative approach, using the architecture shown in Figure 2. The basic idea is to predict the difference between stimulus- and system-level scores rather than predicting directly at the stimulus level, thus leveraging the robustness of the system-level predictions.

3.2. Features and systems evaluated

Referring to previous works [15, 16], stimulus-level acoustic features were built from 40 P.563 internal features [2], 13 mel-frequency cepstral coefficients (MFCCs) including the 0^{th} coefficient, and logarithmic F_0 ($\log F_0$) with first and second derivatives. 12 modified group delay cepstral features (MGD) [23] with parameters $\alpha = 0.6$ and $\gamma = 1.3$ were also included, since these proved successful at detecting synthetic speech in [24]. The signals were downsampled to 8 kHz for P.563 and to 16 kHz for MFCC and $\log F_0$ parameter extraction. MFCCs, $\log F_0$, and MGD were extracted every 10 ms and their means and standard deviations were used in the stimulus-level feature vector, bringing its total dimensionality to 96. The system-level feature vector was composed of the mean, median, and maximum of the stimulus-level acoustic features, plus two binary features: whether the speaker was male or female, and whether the synthesizer was based on waveform concatenation or not.

Each BC uses a different corpus. To reduce the impact of acoustic differences between corpora, per-year mean normalization and per-system variance normalization were applied for system- and stimulus-level prediction, respectively. This normalization scheme improved performance the most.

Two basic regression techniques were evaluated using leave-one-year-out cross-validation:

LR: Elastic net linear regression [25] where the hyperparameter α was 0.005 and the regularization weights were determined by 5-fold cross-validation.

NN: A feed-forward neural network. For system- and stimulus-level prediction, one hidden layer having 64 units and two hidden layers having 16 units were used for the structure, respectively.

Several variants of hierarchical prediction {**LR**, **NN**}+{**LR**, **NN**}, where the former and the latter denote system- and stimulus-level prediction, respectively, were evaluated along with a simple baseline **LR** shown in Figure 1. In **NN+NN**, both predictors were optimized simultaneously through back-propagation. **NN+LR** was excluded due to the small amount of training data: the total number of TTS systems was only 85.

3.3. Experimental results

Table 3 shows the RMSE and ρ_s between observed and predicted MOS. **LR+LR** significantly outperformed **LR** in both prediction tasks. This indicates that system-level scores can be appropriately predicted from system-level feature vectors, rather than by averaging predicted stimulus-level scores. Moreover, the hierarchical structure appears to improve stimulus-level prediction because of correlation between the two levels. Comparing **LR+LR** and **LR+NN**, there was almost no difference in accuracy, possibly because the mean and the SD of frame-level acoustic features cannot capture local degradations. **NN+NN** obtained the lowest system-level RMSE, indicating that NNs may be helpful in predicting overall synthesizer performance. Simultaneous optimization might compensate for the lack of training material.

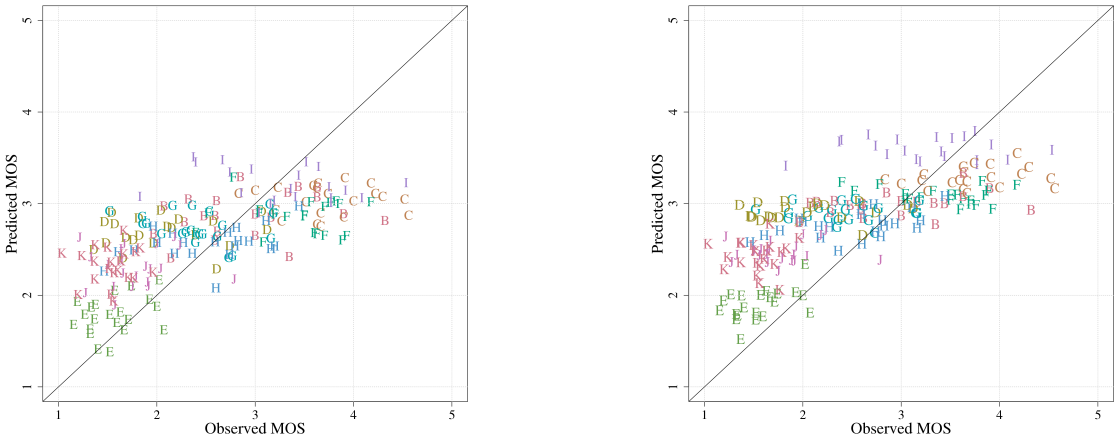
4. Speech naturalness prediction by CNNs

To capture both local and global degradations, it appears essential to consider the feature sequence itself rather than summary statistics, as discussed in Section 3.3. Whereas standard feed-forward NNs cannot handle variable-length inputs in a position-invariant way, CNNs were developed to solve exactly this problem. A CNN is composed of multiple feature-extraction stages. Each stage consists of a convolutional layer, followed by a non-linear transformation and a pooling (subsampling) layer. Due to their hierarchical structure, CNNs may automatically capture degradations at different levels in a variable-length input sequence. Multiple different pooling operations at each stage may help detect different kinds of degradation.

Figure 3 illustrates the structure of a CNN as considered in the experiments, but with one convolutional-pooling layer. A number of feature maps are extracted from the input feature sequence through convolution filters. In the last pooling layer, complete time-invariance is introduced by performing a global pooling operation across time. The pooled features are appended to standard stimulus-level features as in Section 3.2 and then fed into the regression layer for predicting the stimulus-level score.

Table 3: *RMSE and Spearman's rank correlation coefficient between held-out listener MOS and predicted MOS.*

Level	Measure	**LR**	**LR+LR**	**LR+NN**	**NN+NN**
System	RMSE	0.52	0.43	0.43	0.33
	ρ_s	0.55	0.74	0.74	0.72
Stimulus	RMSE	0.78	0.68	0.68	0.68
	ρ_s	0.40	0.56	0.57	0.57



(a) **LR+LR** (b) **LR+CNN**

Figure 4: *Scatter plots of observed and predicted MOS on BC 2012 data, with letters denoting different TTS systems: in the left and right plots, the overall stimulus-level ρ_s are 0.73 and 0.79 while the average within-system stimulus-level ρ_s are 0.04 and 0.18, respectively.*

4.1. Experimental setup

CNN-based approaches were evaluated similarly to the experiment in Section 3.2. Specifically, a 10-second long mean- and variance-normalized MFCC sequence was provided as input to the following network:

CNN: A two-layer convolutional neural network. 1st convolution with 4 filters of size 13×15; 1st max-pooling over 2 time units; 2nd convolution with 26 filters of size 1×15; 2nd max-pooling over all time units. To focus on the effects of convolution and pooling, no fully-connected hidden layers were used, so that utterance-level features including P.563 internal features were fed directly to the output layer.

4.2. Experimental results

The experimental results are shown in Table 4. Overall, stimulus-level **CNN** performance was similar to **LR**. While the training error was found to decrease substantially compared to **LR**, the test-set error did not improve, despite using normalization and regularization techniques such as dropout [26]. It may be that acoustic variation due to different linguistic contexts, speaker characteristics, speaking-style, TTS systems, etc., makes it difficult to construct a feature extractor that works well on arbitrary input sequences, at least from the relatively limited data available here. Providing additional side information, *e.g.*, time-aligned phone identity and degradation annotation, might go some way towards alleviating this issue.

On the other hand, the final row of Table 4 also shows that CNNs improved the correlation between predicted and observed stimulus MOS within each system. This indicates that CNNs had some success in identifying local signal features affecting human perceptual response. Figure 4 presents scatter plots of BC 2012 predictions, from which it can be seen that within-system correlations are higher for **LR+CNN** than for **LR+LR**.

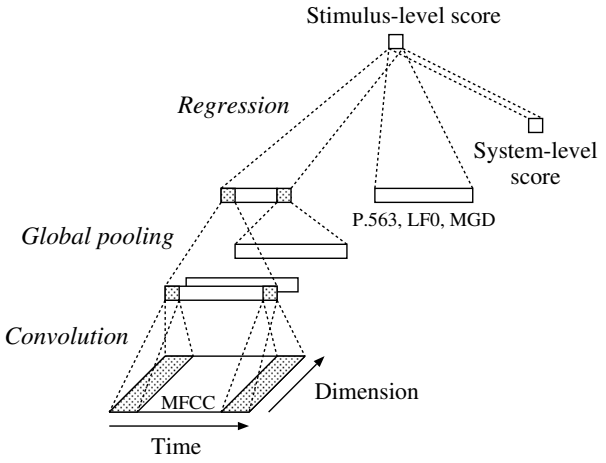


Figure 3: *CNN for stimulus-level speech-quality prediction.*

5. Conclusions

This paper investigated hierarchical and convolutional neural network approaches for speech-quality prediction, specifically naturalness. Despite the limited amounts of training data, neural networks improved several aspects of the predictions. Future work is to augment synthetic-speech acoustic features with linguistic information, and to investigate the utility of objective naturalness predictions for improving TTS systems.

Acknowledgements: The authors thank the Blizzard Challenge organizers for having made the data publicly available. This research was partly funded by Core Research for Evolutionary Science and Technology (CREST) from the Japan Science and Technology Agency (JST), the President's Discretionary Fund of Nagoya Institute of Technology (NIT), EPSRC Programme Grant EP/I031022/1 Natural Speech Technology (NST), and EP/J002526/1 (CAF).

Table 4: *Stimulus-level RMSE and Spearman's rank correlation coefficient between observed and predicted MOS.*

Level	Measure	**LR+LR**	**LR+CNN**
Stimulus	RMSE	0.68	0.69
	ρ_s	0.56	0.58
Stimulus (within-system)	ρ_s	0.11	0.17

6. References

[1] R. F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, vol. 1, pp. 125–128, 1993.

[2] ITU-T Recommendation P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," *ITU-T P-Series Recommendations*, 2004.

[3] T. H. Falk, S. Möller, V. Karaiskos, and S. King, "Improving instrumental quality prediction performance for the Blizzard Challenge," *Proceedings of the Blizzard Challenge workshop 2008*, 2008.

[4] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band networks and speech codes," *ITU-T P-Series Recommendations*, 2001.

[5] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the PESQ measure," *Proceedings of Forum Acusticum 2005*, pp. 2725–2728, 2005.

[6] D.-Y. Huang, "Prediction of perceived sound quality of synthetic speech," *Proceedings of APSIPA ASC 2011*, 2011.

[7] U. Remes, R. Karhila, and M. Kurimo, "Objective evaluation measures for speaker-adaptive HMM-TTS systems," *Proceedings of the 8th ISCA Speech Synthesis Workshop*, pp. 177–181, 2013.

[8] A. W. Black and K. Tokuda, "The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets," *Proceedings of Interspeech 2005*, 2005.

[9] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," *Proceedings of the Blizzard Challenge workshop 2008*, 2008.

[10] S. King and V. Karaiskos, "The Blizzard Challenge 2009," *Proceedings of the Blizzard Challenge workshop 2009*, 2009.

[11] ——, "The Blizzard Challenge 2010," *Proceedings of the Blizzard Challenge workshop 2010*, 2010.

[12] ——, "The Blizzard Challenge 2011," *Proceedings of the Blizzard Challenge workshop 2011*, 2011.

[13] ——, "The Blizzard Challenge 2012," *Proceedings of the Blizzard Challenge workshop 2012*, 2012.

[14] ——, "The Blizzard Challenge 2013," *Proceedings of the Blizzard Challenge workshop 2013*, 2013.

[15] F. Hinterleitner, S. Möller, T. H. Falk, and T. Polzehl, "Comparison of approaches for instrumentally predicting the quality of text-to-speech systems: data from Blizzard Challenges 2008 and 2009," *Proceedings of the Blizzard Challenge workshop 2010*, 2010.

[16] C. R. Norrenbrock, F. Hinterleitner, U. Heute, and S. Möller, "Towards perceptual quality modeling of synthesized audiobooks Blizzard Challenge 2012," *Proceedings of the Blizzard Challenge workshop 2012*, 2012.

[17] W. V. Lukas Latacz, "Double-ended prediction of the naturalness ratings of the Blizzard Challenge 2008-2013," *Proceedings of Interspeech 2015*, pp. 3486–3490, 2015.

[18] C. Mayo, R. A. J. Clark, and S. King, "Listeners' weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis," *Speech Communication*, vol. 53, no. 3, pp. 311–326, 2011.

[19] ITU-R Recommendation BS.1387-1, "Method for objective measurements of perceived audio quality," *ITU-R Recommendations*, 2001.

[20] P. L. Callet, C. Viard-Gaudin, and D. Barba, "A convolutional neural network approach for objective video quality assessment," *IEEE Transactions on Neural Networks*, vol. 17, no. 5, pp. 1316–1327, 2006.

[21] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press LLC, 1993.

[22] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," *Proceedings of ICASSP 2004*, pp. 409–412, 2004.

[23] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," *Proceedings of Interspeech 2012*, pp. 1700–1703, 2012.

[24] Z. Wu, P. L. D. Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z.-H. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, and J. Yamagishi, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 768–783, 2016.

[25] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, vol. 67, no. 2, pp. 301–320, 2005.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.