Non-Intrusive Bit-Rate Detection of Coded Speech

Dushyant Sharma, Uwe Jost
Nuance Communications Inc.
Burlington, USA
Email: dushyant.sharma@nuance.com

Patrick A. Naylor
Electrical & Electronic Engineering
Imperial College London, UK
Email: p.naylor@imperial.ac.uk

***Abstract*—We present a non-intrusive codec type and bit-rate detection algorithm that extracts a number of features from a decoded speech signal and models their statistics using a Deep Neural Network (DNN) classifier. We also present a method for reducing the computational complexity and improving the robustness of the algorithm by pruning features that have a low importance and high computational cost using a CART binary tree. The proposed method is tested on a database that includes additive noise and transcoding as well as a real voicemail database. We show that the proposed method has 25% lower complexity than the baseline, 19% higher accuracy in the bit-rate detection task and 10% higher accuracy in the CODEC classification experiment.**

***Index Terms*—CODEC-Identification, Deep-Neural-Network, Bit-Rate, Voicemail-Classifcation, Speech-Quality.**

I. Introduction

Efficient transmission of speech signals over telephony or VoIP connections typically requires compression using one or more speech codecs. The configuration used for a specific call depends on terminal and network capabilities but also on the network load at the time of the call. The presence of a particular codec has been shown to have adverse effects on many speech processing systems. The type of codec used in the transmission channel has an impact on speech quality [1] and it has been shown that the presence of a GSM codec significantly degrades the performance of speaker identification and verification systems [2]. A four class codec identification algorithm was recently shown to help improve speaker diarization performance [2]. Similarly, identification of the codec can help validate the authenticity of a recording for audio forensics. The effect of codec bit-rate on Automatic Speech Recognition (ASR) has also been widely reported [3], including a study [4] which showed that significantly higher error rates were observed for low bit-rate codecs and tandeming codecs dramatically worsened the recognition. Moreover, the received signal at an Interactive Voice Response (IVR) system is typically a decoded linear PCM signal, with no information about the sequence of codec(s) that were applied. A non-intrusive (without a need for the original unprocessed signal) bit-rate detection algorithm can thus be used to improve a number of speech processing systems as well as being a useful analytics tool for telecommunications traffic.

Two common paradigms in speech coding include waveform coding and analysis-by-synthesis coding [5]. The waveform coders are designed to reproduce the time domain waveform as accurately as possible and the G.711 [6] codec is used in the public switched telephone network and operates at 64 kbps [5]. The analysis-by-synthesis methods are based on a linear prediction model and apply perceptual distortion measures to reproduce only the important characteristics of the signal [5] with examples including the LPC based GSM-FR codec [7] and the CELP [8] based AMR codec [9], which are widely deployed in digital cellular networks.

An algorithm for GSM-FR codec verification is presented in [10], where the spectral properties of the decoded signal are modeled with Gaussian distributions of the quadratic coefficient of a second order polynomial obtained from training data. A more recent study presents a Spectral Harmonic Decomposition (SHD) based codec identification method that uses a correlation based classifier and is able to identify five types of codec with hit rates higher than 92% [11]. The algorithm proposed by Jenner *et al.* [12] extends this approach of a correlation based classifier and noise template based feature extraction. An algorithm for detecting the type of handset used to make a call is presented in [13]. A recent algorithm using deep learning with raw audio for detecting transcoded AMR signals was presented in [14]. The methods for codec identification are mostly tested with clean speech transmission.

In this paper we present a non-intrusive codec bit-rate detection algorithm with applications in ASR and data analytics. The feature extraction for this algorithm is based on our previous work [15]. In this paper we novelly exploit feed-forward DNNs to model the features and also present a novel approach for reducing the computational complexity of the DNN classifier with feature subset selection using Classification And Regression Trees (CART) [16]. The proposed method is shown to reliably estimate the bit-rate of a transcoded speech signal in additive noise conditions down to 10 dB SNR. Furthermore, we present results for classifying real voicemail data using our approach. The remainder of this paper is organized as follows. In Section 2 we present the baseline and proposed algorithms. The databases and evaluation metrics are outlined in Section 3. The results are presented in Section 4 and finally, conclusions are drawn in Section 5.

II. Algorithms

A. NICO-B

The Non-Intrusive CODEC Baseline (NICO-B) algorithm [15] is a data driven algorithm for detecting the type and bit-rate of codec from a speech signal and to the

ISBN 978-0-9928626-7-1 © EURASIP 2017

best of the author's knowledge, this is the only method for codec classification whose performance has been published on noisy, narrow-band telephony data. Therefore, NICO-B is included here as a baseline algorithm. The method begins with short-time segmentation of the decoded speech signal (linear PCM) into 20 ms non-overlapping frames from which an 82 dimensional feature vector is extracted for each frame. The 82 per frame features are characterized by their statistical descriptors in the form of mean, variance, skewness and kurtosis. Additionally, 16 features characterizing the long-term spectral deviation are also included, resulting in 344 global features which are used to train a CART classification tree along with the class labels for the training data. No Voice Activity Detector (VAD) is used in the feature statistics calculations for NICO-B and experiments confirmed that the best results were obtained when feature statistics were computed over all frames in the signal.

B. NICO-FR

The NICO Feature Reduced (NICO-FR) algorithm has a reduced computational complexity and improved robustness (particularly generalization performance) compared to the NICO-B algorithm. This is achieved by pruning some of the features using a CART based feature extraction (discarding features with low importance and high computational complexity). Also, the CART classifier in NICO-B is replaced by a Deep Neural Network (DNN) classifier. The pitch, iSNR and the 16 long term spectral deviation based global features are removed from the NICO-B feature set using the complexity control described below. This results in a 25% lower relative Real Time Factor (RTF), relative to using the full NICO-B feature set. An overview of the NICO-FR algorithm is presented in Fig. 1. The left side of the figure shows test phase where the feature extraction is followed by evaluating the DNN. In the training phase, the CART analysis is carried out to identify feature importance and along with feature complexity, a pruning decision is made and is followed by DNN training. The Power spectrum of Long term Deviation (PLD) flatness and Hilbert envelope features were found to be important for the bit-rate detection tasks and are described in more detail in the following subsections.

1) Hilbert envelope: The Hilbert decomposition of a signal results in a rapidly varying fine structure component and a slowly varying envelope, which has been shown to be an important factor in speech perception [17]. The envelope for frame i of the decoded speech signal $y(i)$ is calculated as:

$$e(i) = \sqrt{y(i)^2 + \mathscr{H}(y(i))^2}, \tag{1}$$

where $\mathscr{H}\{.\}$ is the Hilbert transform. The variance, $\sigma_{e(i)}$ and dynamic range, $\Delta_{e(i)}$ of the envelope for each of the N_i frames are computed as:

$$\sigma_{e(i)} = \frac{1}{N_i}\sum_{i=1}^{N_i}(e(i) - \mu_{e(i)})^2 \tag{2}$$

$$\Delta_{e(i)} = |\max(e(i)) - \min(e(i))|. \tag{3}$$

2) PLD Flatness: The Long Term Average Speech Spectrum (LTASS) [18], is a model for long term shape of the frequency magnitude of a clean speech spectrum and has been used in a number of speech processing algorithms, such as blind channel identification [19]. The Power spectrum of Long term Deviation (PLD) feature for frame i and frequency bin k is defined as:

$$\text{PLD}(i,k) = \log(P_y(i,k)) - \log(P_{LTASS}(k)), \tag{4}$$

where $P_y(i,k)$ is the magnitude power spectrum of noisy signal and $P_{LTASS}(k)$ is the LTASS power spectrum. This measures the effects on the frequency magnitude spectrum caused by distortions and the per-frame LTASS deviation spectrum is used to derive the spectral flatness (SF) features as follows:

$$SF(i) = \frac{\exp\left(\frac{1}{N_k}\sum_{k=1}^{N_k}\log(\text{PLD}(i,k))\right)}{\frac{1}{N_k}\sum_{k=1}^{N_k}\text{PLD}(i,k)}, \tag{5}$$

where k is the FFT bin index and N_k is the number of FFT bins. The PLD spectral flatness and its rate of change over subsequent frames are included as short-term features.

3) Complexity Controlled Classifier: A number of DNN architectures have been proposed in the literature and are considered the state of the art machine learning algorithms in a number of applications, including automatic speech recognition [20]. A DNN is a feed-forward artificial neural network with a number of non-linear hidden units connected between an input and output layer. The nodes in each layer are connected with nodes in adjacent layers and each connection is scaled by a coefficient. The nodes are modelled with a non-linear activation function, in our case we use the sigmoid function. The output layer for a multi-class classification problem typically uses the softmax function [20]. A strong advantage of DNNs is that they can be discriminatingly trained by back-propagating the derivatives of a cost function that measures the difference between the desired and estimated output and adjusting the weights of the network in a fine-tuning stage using for example the Low memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) optimization algorithm [21].

A common issue with DNNs is the lack of an efficient feature subset selection algorithm [22] that can help control the complexity of the system. For this purpose we propose the use of a CART [16] binary tree algorithm as a method to determine features with a low importance to the classification task, independent from the DNN structure. A CART classification tree is constructed using the training data, using the deviance split criterion (a negative log likelihood) to grow an initial tree. The tree is then pruned to a reduced size using 10 fold cross validation. The feature significance can then be computed by summing the change in deviance caused by splits in the final pruned model for each feature and dividing by the corresponding number of branch nodes [16]. This in combination with the computational complexity of the features forms the basis of the decision to retain or prune away features.

ISBN 978-0-9928626-7-1 © EURASIP 2017

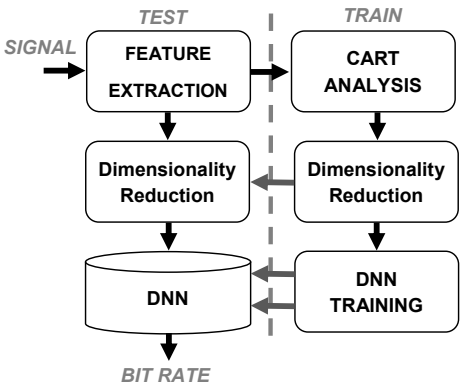

Figure 1. NICO-FR block diagram with the CART based feature analysis in training phase and a DNN classifier in the test phase.

We wish to determine a feature subset $\tilde{\mathbf{\Phi}}$ by minimising a function, $\Omega(\mathbf{\Phi}, \boldsymbol{\tau})$, where $\mathbf{\Phi}$ is a vector of feature importance from CART and $\boldsymbol{\tau}$ is a vector of feature complexity, subject to $\epsilon(\tilde{\mathbf{\Phi}}) = \epsilon(\mathbf{\Phi}) + \delta$, where δ is a tolerance on the overall hit rate and ϵ is the hit rate. In this work we perform this minimization experimentally with $\delta = 0.1\%$ and find that the iSNR, pitch and the 16 PLD based global features could be successfully removed and therefore led to a reduction in the processing time of the algorithm by 25% compared to using the entire feature set. Additional computational and memory reduction was achieved for the DNN model by the reduction in the size of the input layer and therefore fewer weights and bias parameters in the model. The final feature set has dimension of 312 and is used to train a 2 hidden layer DNN classifier (more details are presented in Section.IV).

III. Evaluation

Due to a lack of coded speech database with known codec(s) for the task of codec bit-rate detection, we use the synthesis of codec speech using a database of clean speech and adding noise and codec pairs, described in Section 3.1, following the methodology of other published work in this area. However, in a real telephony system, the speech signal is not only impacted by codecs but also potentially by other detrimental effects like jitter or dropped frames and in order to establish the extent to which the proposed method performs on real data, we created a second database using real voicemail messages from an internal demonstration platform in the UK. Although information about the specific codec bit-rate was not available, it was possible to classify whether the message originated from a landline or mobile phone. This database is described in more detail in Section 3.2.

A. WSJ1C Database

The speech material is taken from the spontaneous partition of the Wall St Journal database(WSJ1) [23] that was designed to facilitate the development and evaluation of large vocabulary, speaker-independent, continuous speech recognition systems. The spontaneous partition includes dictation of 400 sentences spoken by 40 journalists with varying degrees of experience in dictation in US English, split evenly into training and test partitions without any overlap of speech or speaker. For the WSJ1C database, car and babble noises were added to the speech at 10, 20 and 30 dB SNR, with randomized noise segments (to ensure different sections of the noise files are used). A narrowband telephone channel filter was applied prior to the twelve combinations of codecs presented in Table I. This results in 288,000 utterances (20 speakers$\times$200 utterances$\times$2 noises$\times$3 SNRs$\times$12 CODEC combinations) in each of training and test partitions and different noise sources were used in the partitions (no overlap of noise sources in test and train).

CODEC 1		CODEC 2		
Type	**BR(kbps)**	**Type**	**BR(kbps)**	**MBR(kbps)**
G.711	64.0	G.711	64.0	64.0
AMR	12.2	G.711	64.0	12.0
AMR	7.4	G.711	64.0	8.0
AMR	4.75	G.711	64.0	4.75
G.729	8.0	G.711	64.0	8.0
LPCM	128.0	G.711	64.0	64.0
G.711	64.0	GSM-FR	12.0	12.0
AMR	12.2	GSM-FR	12.0	12.0
AMR	7.4	GSM-FR	12.0	8.0
AMR	4.75	GSM-FR	12.0	4.75
G.729	8.0	GSM-FR	12.0	8.0
LPCM	128.0	GSM-FR	12.0	12.0

Table I
THE TWELVE CODEC TRANSCODING PAIRS USED IN THE WSJ1C DATABASE. THE MAIN CODEC'S ARE LINEAR PCM (LPCM), G.711 A-LAW [6], AMR [9], G.729 [24] AND GSM-FR [7].

B. VM Database

The voicemail (VM) database consists of real voicemail messages deposited by employees over a one year period and in demonstrations. The data collection passed through the UK telephone infrastructure and was subject to typical degradations and signalling protocols. The following two cases can be identified in the set of messages:

1) Landline-originated-call (LOC) : this should most likely be a G.711 codec with a 64 kbps bit-rate
2) Mobile-originated-call (MOC): this will most likely be one or more of the GSM/AMR Codec's at bit-rates in the 4.75 kbps to 12.2 kbps range.

The VM database was constructed by randomly selecting 10,000 messages from each of the two classes in the training partition and 1000 messages in each class were assigned to the test partition.

C. Metrics

In addition to the confusion matrix for each experiment we compute the hit rate in each class, defined as the percentage

ISBN 978-0-9928626-7-1 © EURASIP 2017

MBR	PESQ	PESQR (%)	WAR (%)	WARR (%)
64 kbps	3.75	0.0	56.8	0.0
12 kbps	3.45	-8.0	54.1	-5.0
8 kbps	3.41	-9.1	52.8	-7.6
4.75 kbps	3.15	-16.0	49.3	-15.2

Table II
DEGRADATION IN PERCEPTUAL SPEECH QUALITY (PESQ) AND WORD ACCURACY RATE (WAR) RELATIVE TO THE 64 KBPS MINIMUM BIT-RATE (MBR) CONDITION IN THE WSJ1C DATABASE. THE WAR WAS OBTAINED USING A KALDI RECOGNIZER.

of utterances correctly classified, as follows,

$$\text{HR}(\%) = \frac{\sum_{n=1}^{N} \Upsilon(\tilde{\theta}_n, \theta_n)}{N} \times 100, \tag{6}$$

where $\tilde{\theta}_n$ is the estimated class label according to some detection criteria (i.e. codec bit-rate detection) and θ_n is the actual class label for the n^{th} speech utterance. The total number of utterances in the test set is N and $\Upsilon(a, b)$ is defined as:

$$\Upsilon(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

IV. RESULTS

Here we present the results for the two classification tasks. The DNN architecture selected using a grid search for the bit-rate classification task on the WSJ1C database was 312x90x60x4 and 312x90x60x2 for the VM database.

In Table II we present the mean and relative degradation in PESQ [25] and word accuracy rate (WAR) for the Kaldi ASR system [26] trained on the WSJ1 database (linear PCM without a speech CODEC or additive noise). The Kaldi system was based on a GMM-HMM acoustic model trained with MFCC features transformed using LDA-MLLT and a trigram language model. The final hypothesis is decoded using minimum Bayesian-Risk re-scoring. It can be seen that as the bit-rate is reduced from 64 kbps to 4.75 kbps, a 16.0% and 15.2% relative reduction is seen in PESQ and WAR respectively, showing that the codec bit-rate has a significant impact on perceptual speech quality and ASR performance.

The performance of the NICO-B and NICO-FR algorithms for the bit-rate detection task on the WSJ1C database is presented in Table III and Table IV, with overall hit rates being 76.4% and 95.4% respectively. The proposed NICO-FR algorithm has a higher overall hit rate and better classification in all the 4 classes than NICO-B. A similar pattern is seen on the VM database where the task was to identify the origin type of real voicemail messages. The confusion matrix for the NICO-B and NICO-FR algorithms for CODEC classification on the VM database are presented in Table V and Table VI respectively. Also shown are the overall hit rates, which are 84.5% for NICO-B and 94.5% for the NICO-FR algorithm. Again, NICO-FR is shown to have a 10% higher overall hit rate for this task.

	NICO-B				
	Predicted				
Actual	**64 kbps**	**12 kbps**	**8 kbps**	**4.75 kbps**	**HR (%)**
64 kbps	37524	183	406	287	97.7
12 kbps	2715	50397	11431	12256	65.6
8 kbps	6742	22044	63949	22465	55.5
4.75 kbps	501	2152	5039	49908	86.6
			Mean HR(%)		76.4

Table III
CONFUSION MATRIX FOR BIT-RATE DETECTION ON THE TEST PARTITION OF THE WSJ1C DATABASE FOR NICO-B. EACH ELEMENT OF THE MATRIX REPRESENTS THE NUMBER OF UTTERANCES. THE HR COLUMN PRESENTS THE HIT RATE FOR EACH CLASS.

	NICO-FR				
	Predicted				
Actual	**64 kbps**	**12 kbps**	**8 kbps**	**4.75 kbps**	**HR (%)**
64 kbps	38072	188	138	3	99.1
12 kbps	72	74684	1997	47	97.2
8 kbps	2167	5052	102003	5977	88.5
4.75 kbps	16	341	1538	55705	96.7
			Mean HR (%)		95.4

Table IV
CONFUSION MATRIX BIT-RATE DETECTION ON THE TEST PARTITION OF THE WSJ1C DATABASE FOR NICO-FR. EACH ELEMENT OF THE MATRIX REPRESENTS THE NUMBER OF UTTERANCES. THE HR COLUMN PRESENTS THE HIT RATE FOR EACH CLASS.

	NICO-B		
	Predicted		
Actual	**LOC**	**MOC**	**HR (%)**
LOC	845	154	84.6
MOC	156	843	84.4
	Mean HR(%)		84.5

Table V
CONFUSION MATRIX FOR DETECTING ORIGINATING CODEC ON VM DATABASE FOR NICO-B. EACH ELEMENT OF THE MATRIX REPRESENTS THE NUMBER OF UTTERANCES. THE HR COLUMN PRESENTS THE HIT RATE FOR EACH FOR EACH CLASS.

V. CONCLUSIONS

We presented a novel application of a non-intrusive codec bit-rate detection algorithm based on speech feature statistics modelled with a complexity-controlled feed-forward DNN. The CART based complexity control helped achieve a 25% reduction in computational complexity without loss in classification error rate. This is important for application in a real world application as the computational complexity is directly related to the cost of providing a service and even small reductions in compute requirements can build up when

ISBN 978-0-9928626-7-1 © EURASIP 2017

NICO-FR			
	Predicted		
Actual	**LOC**	**MOC**	**HR (%)**
LOC	932	67	93.3
MOC	43	956	95.7
	Mean HR(%)		94.5

Table VI
CONFUSION MATRIX FOR DETECTING ORIGINATING CODEC ON VM DATABASE FOR NICO-FR. EACH ELEMENT OF THE MATRIX REPRESENTS THE NUMBER OF UTTERANCES. THE HR COLUMN PRESENTS THE HIT RATE FOR EACH CLASS.

processing a large number of signals, such as in a telecommunications infrastructure. We presented two experiments, first to identify one of four bit-rates of transcoded speech with two noises added at 10, 20 and 30 dB SNR. In this task the proposed NICO-FR algorithm achieved a mean hit rate of 95.4% compared to 76.4% with the baseline algorithm. In a second experiment, the NICO-FR algorithm was trained to detect the origin of a voicemail message using phone number meta data as being a land-line originated or mobile originated call. In this task our approach gave a mean hit rate of 94.5%, validating the potential for deploying this method for classifying real telephony data. Moreover, the NICO-FR method was shown to be robust to additive noise with performance comparable to other studies on codec identification on clean speech.

REFERENCES

[1] T. Lugwig and U. Heute, "Detection of digital transmission systems for voice quality measurements," in *Proc. European Conf. on Speech Communication and Technology*, 2001, pp. 1699–1702.

[2] R. Weychan, A. Stankiewicz, T. Marciniak, and A. Dabrowski, "Improving of speaker identification from mobile telephone calls," *in Trans. Multimedia Communications, Services and Security*, 2014.

[3] B. Lilly, K. K. Paliwal *et al.*, "Effect of speech coders on speech recognition performance," in *Proc. Intl. Conf. on Spoken Lang. Processing (ICSLP)*, vol. 4. IEEE, 1996.

[4] A. Gallardo-Antolin, C. Pelaez-Moreno, and F. Diaz-de Maria, "Recognizing gsm digital speech," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1186–1205, Nov 2005.

[5] J. Gibson, "Speech coding methods, standards, and applications," *IEEE Circuits Syst. Mag.*, vol. 5, no. 4, pp. 30–49, 2005.

[6] ITU-T, *Pulse Code Modulation (PCM) of Voice Frequencies*, International Telecommunications Union (ITU-T) Rec. G.711, Nov. 1998.

[7] E. T. S. I. (ETSI), *GSM 06.10: Full Rate (FR) Speech Transcoding*, European Telecommunications Standards Institute (ETSI) Recommendation GSM 6.10, 1995.

[8] M. Schroeder and B. Atal, "Code-excited linear prediction(CELP): High-quality speech at very low bit rates," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 10, 1985, pp. 937–940.

[9] *GSM 06.90: Adaptive Multi-Rate (AMR) Speech Transcoding*, European Telecommunications Standards Institute (ETSI) Recommendation GSM 06.90, 1998.

[10] T. Lugwig, "Comfort noise detection and GSM-FR-CODEC detection for speech-quality evaluation in telephone networks," in *Proc. Intl. Conf. on Spoken Lang. Processing (ICSLP)*, 2002, pp. 309–312.

[11] K. Sholz, L. Leutelt, and U. Heute, "Speech-codec detection by spectral harmonic-plus-noise decomposition," in *Proc. Asilomar Conference on Signals, Systems and Computers*, 2004, pp. 2295–2299.

[12] F. Jenner and A. Kwasinski, "Highly accurate non-intrusive speech forensics for CODEC identification from observed decoded signals," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, March 2012.

[13] C. Kotropoulos, "Source phone identification using sketches of features," *in IET Biometrics*, vol. 3, no. 2, pp. 75–83, 2014.

[14] D. Luo, R. Yang, and J. Huang, "Detecting double compressed amr audio using deep learning," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.

[15] D. Sharma, P. A. Naylor, N. D. Gaubitch, and M. Brookes, "Non intrusive CODEC detection algorithm," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Mar. 2012.

[16] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. CRC Press, 1984.

[17] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Letters to Nature*, vol. 416, pp. 87 – 90, 2002.

[18] ITU-T, *Artificial Voices*, International Telecommunications Union (ITU-T) Recommendation P.50, Sep. 1999.

[19] N. D. Gaubitch, M. Brookes, and P. A. Naylor, "Blind channel identification in speech using the long-term average speech spectrum," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Glasgow, Aug. 2009.

[20] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82 – 91, 2012.

[21] D. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical Programming*, vol. 45, 1989.

[22] Y. Li, C.-Y. Chen, and W. Wasserman, "Deep feature selection: Theory and application to identify enhancers and promoters," in *Journal of Computational Biology*. Springer, 2015.

[23] P. L. D. Consortium. (1994) Csr-ii (wsj1) complete ldc94s13a. DVD. [Online]. Available: https://catalog.ldc.upenn.edu/LDC94S13A

[24] ITU-T, *Coding of Speech at 8 kbit/s using Conjugate-Structure Algebraic-Code-Excited Line-Prediction (CS-ACELP)*, International Telecommunications Union (ITU-T) Recommendation G.729, Mar. 1993.

[25] ——, *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, International Telecommunications Union (ITU-T) Recommendation P.862, Feb. 2001.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

ISBN 978-0-9928626-7-1 © EURASIP 2017