Investigation of Deep Neural Networks for Robust Recognition of Nonlinearly Distorted Speech

Ladislav Seps, Jiri Malek, Petr Cerva and Jan Nouza

Faculty of Mechatronics, Informatics, and Interdisciplinary Studies,
Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic.
ladislav.seps@tul.cz

Abstract

This paper studies the use of hybrid context-dependent Deep Neural Network Hidden Markov Model (DNN-HMM) architecture for robust recognition of speech affected by real-world nonlinear distortions. We consider two types of distortions; a) signals distorted through overgained microphone preamplifier in the analog domain and b) recordings exhibiting unnatural spectral sparseness, caused by excessive denoising or low-bit-rate compression. We compare the performance of DNN-HMM architecture with that of the conventional system, based on context-dependent Gaussian Mixture Model (GMM)-HMMs, which applies channel/speaker adaptation and/or feature compensation in the front-end via Histogram Equalization (HEQ). We show that DNN-HMM architecture achieves a significantly lower Word Error Rate (WER) on the considered distorted datasets and that the obtained relative WER reduction is higher than 60%. We also investigate the usefulness of the feature compensation via HEQ for a DNN-HMM system and show that it can be helpful in the case of shallower networks.

Index Terms: Deep Neural Networks, Real-world Nonlinear Distortion, Robust Speech Recognition.

1. Introduction

A new type of acoustic model, based on hybrid DNN-HMM architecture [1], has been introduced recently. These models yield significant WER reduction compared with conventional speech recognition systems based on GMM-HMMs in various tasks, e.g., context-independent phoneme recognition [2], large vocabulary speech recognition [1] and multilingual training [3].

This success motivates the utilization of DNN-HMMs for recognition of speech distorted by environmental conditions, such as additive noise or convolutive channel distortion. The work [4] demonstrates robustness of DNN-HMMs in a medium vocabulary task from the Aurora 4 noise database [5], which is based on a Wall Street Journal corpus. The work [6] investigates the utilization of speech enhancement preprocessing techniques, usually used to improve the noise-robustness of GMM-HMM recognizers, as preprocessing for DNN-HMMs systems.

The distortions considered above can be described via classical linear models. In this paper, we investigate the robustness of DNN with respect to specific types of other real-world distortions, for which the linearity of the distorting transformation does not hold. First, we focus on a distortion caused by nonlinear amplification (or even clipping) in the analog signal domain, which is followed by sampling and coding via a lossy codec optimized for speech perceptual quality (in our case Windows Media Audio Codec 2). For simple reference, we will denote this distortion as NAD (Nonlinear Amplification Distortion). Second, we consider speech recordings that exhibit unnatural spectral sparsity. This effect originates from a lossy compression to a very low bit-rate. Similar problems are exhibited by signals processed via denoising algorithms based on spectral subtraction. These techniques set to zero frequency bins that are identified as noise. We will denote this distortion as SMD (Spectral Masking Distortion) below.

In our work, we compare the performance of DNN-HMM architecture achieved on distorted datasets with the conventional GMM-HMM system. However, the performance of the standard GMM-HMM system deteriorates markedly when processing distorted datasets, as we show in the experiments and in [7]. To compensate, we employ two robust speech recognition (RSR) techniques, which are very efficient for the considered distortions: 1) *Histogram Equalization* (HEQ, [8]) is a front-end feature preprocessing technique which is able to invert nonlinear memoryless transformations and exhibits small computational demands; and 2) *Constrained Maximum Likelihood Linear Regression* (CMLLR, [9]) is a channel/speaker adaptation technique applied in the feature domain. It is utilized to compensate for an unknown linear filtering applied by the environment to the utterances. The CMLLR algorithm is computationally demanding and relies on the availability of a reference phonetic transcription (i.e., it requires a two-pass data recognition).

We show that the DNN-HMM system is able to achieve higher recognition accuracy in comparison with GMM-HMM (complemented by considered RSR techniques), especially for datasets affected by unnatural spectral sparseness. Due to its positive influence on GMM-HMM system performance, we examine the usefulness of feature compensation via HEQ in the front-end of a DNN-HMM system. We show that benefits of this preprocessing are limited to shallower networks only.

2. Considered distortions

In our study, we consider the two following types of nonlinear distortions, which affect three datasets mentioned later.

2.1. Nonlinear Amplification Distortion

The NAD distortion is caused by an erroneous excessive setting of the analog preamplifier. The preamplifier becomes saturated and amplifies the input signal in a nonlinear way. In an extreme case, the signal may become clipped prior to sampling. Subsequently, the signal is sampled and coded by a lossy codec. After coding, the potential clipping becomes difficult to detect, as the characteristic flat amplitude level disappears in the signal domain (see Figure 1 for example).

The nonlinear amplification and the potential clipping af-

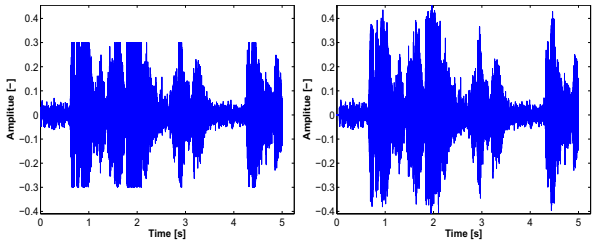

Figure 1: Uncompressed speech signal affected by clipping (left) and the same decompressed speech signal after lossy coding into wma2 (44.1 kHz, 80 kbps, right).

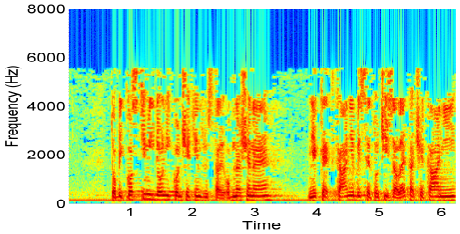

Figure 2: The spectrogram of an mp3 encoded utterance at the bit-rate of 16 bit/s.

fect perceptual quality of the speech [10] and deteriorate the accuracy of ASR, as will be shown later in this paper. In the spectral domain, the nonlinear transformation causes spectral spreading, i.e., new harmonic frequencies appear in the output signal [11]. This is unlike linear filtering, where harmonic components are amplified/delayed, but no new components appear.

Our dataset distorted by NAD consists of eight *lectures* given in Czech (11 hours and 45 minutes of speech, 85396 words), recorded for streaming purposes at our university. The signal is captured by a close-talk microphone. The common background noise of a lecture hall is present in the recording. The recordings were originally sampled at 44.1 kHz and then compressed by wma2 lossy codec (constant bit-rate 48 kbps), optimized for perceptual quality suitable for human listeners. Prior to recognition, the signals were downsampled to 16 kHz.

2.2. Spectral Masking Distortion

Spectral Masking Distortion arises when certain spectral components are, based on their magnitude, removed from the speech spectra. We distinguish two cases: 1) a very low-bit-rate compression (in our case to mp3 format) or 2) excessive spectral subtraction denoising.

Low-bit-rate *mp3 compression* neglects frequency components which are considered inaudible based on a psychoacoustic model. Thus, the reconstructed (decompressed) signal exhibits many zeros in the time-frequency domain. The compression to low bit-rates (<24 kbit/s) causes suppression of phonemes at word boundaries, which deteriorates the ASR accuracy [12].

We consider a 16kbit/s bit-rate at a 16 kHz sampling frequency. An example spectrogram of a compressed utterance is shown in Figure 2. In this case, all frequencies above 5.5 kHz and selected frequencies in the band 0–5.5 kHz are discarded.

The dataset distorted by SMD caused by low-bit-rate mp3 compression consists of 22 recordings (1 hour and 12 minutes of speech, 8096 words) of *radio broadcasts*. Spontaneous speech by various speakers was recorded at a sample rate of 16 kHz.

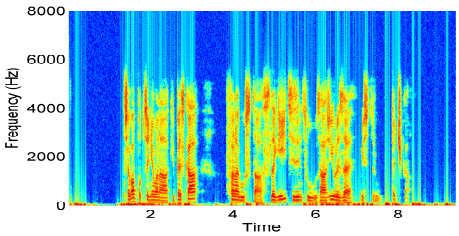

Figure 3: Spectrogram of an utterance distorted by SMD caused by a denoising method.

Subsequently, an mp3 compression via LAME encoder was applied at a constant bit-rate of 16 kbit/s in order to present the recordings on the web page of a radio station.

The *denoising* based on spectral subtraction is designed to remove stationary noise from signals. A power spectrum of noise is estimated and subsequently subtracted from spectrum of speech. This results in removal of low-magnitude spectral components, which are assumed to come from the noise. A spectrogram of a processed utterance is shown in Figure 3. It reveals the suppression of some harmonics and of all frequencies above 5.5 kHz. The impact of the suppression is the most apparent in the gaps between words, where almost all frequencies are put equal to zero.

The dataset affected by excessive denoising consists of 1161 *short Czech utterances* read by various speakers, recorded via a close-talk microphone. The total duration of the dataset is 1 hour and 45 minutes, and it contains 12780 words. The original sampling frequency of 44.1 kHz was downsampled to 16 kHz. During the recording, a denoising method for removal of stationary noise, provided by software drivers of the sound device, was turned on. No additional compression was applied.

3. Employed Recognition Engine and Investigated Methods

We use our own ASR system. Its core is formed by a one-pass speech decoder performing a time-synchronous Viterbi search.

The linguistic part of the system consists of a lexicon and a language model. The lexicon contains 550k entries (word forms and multi-word collocations) that were observed most frequently in a 10 GB large corpus covering newspaper texts and broadcast program transcripts. Some lexical entries have multiple pronunciation variants. Their total number is 580k.

The employed Language Model (LM) is based on N-grams. For practical reasons (mainly with respect to the very large vocabulary size), the system uses bigrams. In the training word corpus, 159 million unique word-pairs (1062 million in total) belonging to the items in the 550k lexicon were observed. However, 20 percent of all "word-pairs" actually include sequences containing three or more words, as the lexicon contains 4k multi-word collocations. The unseen bigrams are backed-off by the Kneser-Ney smoothing technique [13].

3.1. Baseline GMM-HMM architecture

The baseline acoustic model is speaker independent (SI), it is trained on a database containing 300 hours of undistorted speech recordings based on tied-state context-dependent HMMs of Czech phonemes and several types of non-speech events (e.g., breathing, various hesitation sounds, cough, lip-smack, etc.). The model contains 4k physical states with up to 32

Gaussian components per state (i.e., 120k components in total). Its training proceeds through standard maximum likelihood training procedure followed by discriminative training using the Minimum Phone Error (MPE) criteria [14].

Frames of length 25 ms with frame shift 10 ms weighted by Hamming window are used for feature extraction. The features are 13-dimensional MFCCs with Δ and $\Delta\Delta$. The sampling frequency used is 16 kHz.

3.2. Hybrid DNN-HMM architecture

The hybrid DNN-HMM acoustic model inherits the structure of the baseline GMM-HMM architecture: the DNN is trained to provide scaled likelihood estimates for 4k physical states of the GMM-HMM model. The feature vectors for DNN consist of 11 concatenated MFCC vectors, five preceding and five following the current frame.

For DNN training, Theano library [15] was utilized and we performed so called *discriminative pretraining* [16], where a one-hidden-layer network was trained using 50 epochs of back propagation. We observed no significant improvement in the error rate with more than 50 epochs on our training datasets. Subsequently, the softmax layer was replaced by another randomly initialized hidden layer with a new random softmax layer on top, and this two-hidden-layer DNN was again trained using 50 epochs of back propagation. This process was repeated until the desired number of hidden layers was reached.

Each DNN hidden layer consisted of 1024 units. The mini-batch size was 1000 and the learning rate was 0.08.

3.3. Histogram Equalization

Histogram Equalization [8] is a front-end enhancement technique, frequently utilized by the GMM-HMM systems. It is designed for compensation of memoryless nonlinear distortions and is very efficient for compensation of the distortions considered in this paper. For these reasons, we investigate its potential usefulness for DNN-HMM architecture.

HEQ is applicable within different phases of the feature extraction chain, we utilize it in the MFCC feature domain. HEQ is based on finding a transform such that the cumulative distribution function (CDF) of the transformed feature vectors is the same as the reference CDF, which was derived from the training data in advance. The elements of feature vectors are considered independent, and are processed separately.

We derive the reference CDFs from approximately 30 minutes of recordings that were originally used for the training of the ASR acoustic model. When estimating the CDFs of the distorted feature vectors, all available test data are used. We apply the transform to all MFCC features, including delta and delta-delta features, as was proposed in [17].

3.4. Adaptation to channel/speaker

The adaptation procedure fits the baseline SI acoustic model to a given speaker and/or acoustic channel (i.e., linear filtering, which was applied to the test data). This process is performed in an unsupervised manner in two recognition passes. In the first pass, we utilize the baseline SI model to obtain phonetic transcript of the given recording. The recording is then split into 5-minute-long segments. For each segment, the Constrained Maximum Likelihood Linear Regression (CMLLR) [9] method is employed to estimate a global linear transformation matrix using the created phonetic transcript and the baseline SI GMM-HMM acoustic model. Then, the second speech recognition pass is performed, where the estimated transforms are applied on all feature vectors belonging to individual segments.

4. Experimental results

In this Section, we present a comparison of WER achieved by investigated acoustic modeling techniques on distorted datasets. We consider the following configurations of the recognition process:

- GMM(SI): baseline one-pass speaker-independent (SI) transcription using GMM-HMM architecture.
- DNNx(SI): one-pass SI transcription using DNN-HMM acoustic models with x hidden layers.
- GMM+HEQ(SI): one-pass SI recognition via GMM-HMM architecture with HEQ applied in the front-end.
- DNNx+HEQ(SI): one-pass SI recognition via DNN-HMM architecture with HEQ applied in the front-end.
- GMM(SA): two-pass speaker (channel) adapted (SA) transcription using GMM-HMM acoustic models.
- GMM+HEQ(SA): two-pass SA transcription via GMM-HMM architecture and with HEQ applied in the front-end.

We state the results for networks with $x \in \{2, 3, 5\}$. Our experiments on distorted datasets show that additional layers above the number five do not improve the performance significantly.

4.1. Transcription of the lectures distorted by NAD

A typical WER of our baseline GMM-HMM system on a clean lecture recording is approximately 15%. Unfortunately, the WER of signals distorted by NAD rises dramatically to 36–69%. This broad range stems from the fact that each recording is distorted by NAD up to a certain degree, varying among recordings (the level of preamplification is changed between recording sessions).

The results of our experiments are summarized in Table 1. It can be seen that best results are achieved by GMM+HEQ(SA) and DNN5. Compared to the GMM(SI) baseline, the absolute WER reductions are 5-14% and 6-15%, respectively. This shows that DNN-HMM without any compensation techniques is more robust with respect to NAD compared to GMM-HMM system. However, the performance achieved by DNN-HMMs is still worse compared to the performance yielded by baseline recognizer on undistorted data, which gives motivation for application of feature compensation via HEQ.

The HEQ preprocessing is most beneficial for a single-pass baseline GMM, decreasing the absolute WER by 1-4%. In the case of DNN2 and DNN3, the feature compensation reduces the absolute WER by 0.5-3% and 0-2%, respectively. The lecture M7, which is the least affected by NAD, exhibits slight WER rise (0.4% and 0.2%). The HEQ exhibits similar behavior when applied to a two-pass GMM recognizer. We did not find any benefit of using the HEQ as a front-end for DNN5.

4.2. Transcription of compressed radio broadcasts

A common WER of the baseline single-pass GMM-HMM system on similar radio broadcasts, which contain spontaneous speech but are unaffected by the SMD, reaches 20%. The compression to 16kbit/s causes a severe increase of WER (about 25%), as is shown in Table 2 and in paper [12]. We evaluate the transcription performance over all recordings in the dataset, as the amount of distortion is constant.

Table 1: WER in [%] for the lecture dataset. M or F denotes male or female lecturer, respectively. The best result achieved for each lecture is given in bold. The gender distribution reflects the incidence of lecturers at our university.

Lec. (NAD)	GMM (SI)	DNN2 (SI)	DNN3 (SI)	DNN5 (SI)	GMM +HEQ (SI)	DNN2 +HEQ (SI)	DNN3 +HEQ (SI)	DNN5 +HEQ (SI)	GMM (SA)	GMM +HEQ (SA)
M1	69.2	63.8	61.8	59.1	64.9	61.6	59.8	57.9	58.6	**57.8**
M2	68.8	59.9	57.8	**54.8**	67.0	59.0	57.6	55.5	55.2	55.1
M3	66.1	60.4	57.5	54.9	63.6	58.0	56.1	53.8	52.6	**52.0**
M4	53.8	50.3	47.9	**44.7**	50.8	49.7	47.0	44.9	48.2	48.6
F1	53.4	45.3	41.2	**38.1**	50.2	42.2	39.6	38.9	42.3	40.2
M5	52.9	47.4	45.3	**42.4**	50.5	46.8	44.2	42.5	48.4	46.9
M6	44.3	40.8	39.4	37.0	42.0	39.9	38.9	**36.5**	39.0	38.8
M7	36.3	34.1	33.4	30.2	35.1	34.5	33.6	30.7	**28.1**	28.3

Table 2: WER in [%] for both datasets distorted by SMD. "Radio" and "Short" denote the radio broadcasts and the short utterances dataset, respectively. The best achieved results are given in bold.

Dataset (SMD)	GMM (SI)	DNN2 (SI)	DNN3 (SI)	DNN5 (SI)	GMM +HEQ (SI)	DNN2 +HEQ (SI)	DNN3 +HEQ (SI)	DNN5 +HEQ (SI)	GMM (SA)	GMM +HEQ (SA)
Radio	45.6	20.7	20.1	**19.0**	28.6	21.3	20.8	20.5	20.7	**19.0**
Short	43.9	17.1	16.1	**14.9**	22.6	16.5	15.8	15.4	41.2	21.1

The best results are yielded by two techniques, the GMM+HEQ(SA) and DNN5. The absolute improvement compared to the baseline is 26.6% for both techniques (i.e., the baseline WER is reduced by almost 60% relatively). The results of the DNN5 indicates that DNN-HMM architecture is able to cope with SMD very efficiently and its accuracy is comparable to that of a baseline GMM-HMM recognizer achieved on undistorted data.

For this dataset, the feature compensation by HEQ brings a great benefit to the GMM-HMM performance, reducing the absolute WER by 17% in the SI case. However, this benefit does not apply to any of the considered DNN-HMMs, which appear to be inherently robust to SMD.

4.3. Transcription of excessively denoised utterances

In this experiment, we process the dataset of short utterances described in Section 2.2. Again, the results are given as average performance values over all utterances. Our SI baseline normally achieves about 10% WER when transcribing similar short utterances read in quiet environment, when no spectral subtraction is applied. In the case when an inappropriate denoising is applied, the WER rises by about 34% absolutely.

For this dataset, the DNN5 distinctly outperforms the GMM+HEQ(SA) configuration and confirms the robustness of DNN-HMM architecture to spectrally sparse utterances. The absolute improvement of DNN5 compared to baseline GMM(SI) is 29% (66% relatively) and the two-pass GMM-HMM yields 22.8% WER reduction (52% relatively).

The single-pass baseline GMM-HMM benefits from the HEQ preprocessing, which yields an absolute WER reduction of 21.3%. The usefulness of HEQ for DNN-HMM is negligible, attaining only a 0.6% and 0.3% absolute improvement for DNN2 and DNN3, respectively. The channel adaptation by itself suffers from short duration of the utterances, which provide only a small amount of data for the estimation of the adaptation transformation.

5. Conclusions

We have experimentally studied the hybrid DNN-HMM based architecture with respect to two types of nonlinear distortions.

For utterances distorted by nonlinear amplification in the analog domain; the results of DNN-HMM without any additional enhancement techniques is comparable to the GMM-HMM system, when endowed with HEQ and speaker/channel adaptation. The absolute WER reduction over considered GMM-HMM baseline is 6-15%. The benefit of using an HEQ front-end for DNN-HMM architecture is limited only to shallower networks (two or three hidden layers), providing 0-3% of absolute WER reduction.

The DNN-HMM architecture shows high robustness with respect to unnatural spectral sparseness of signals, which is very harmful to GMM-HMM systems. The values of WER obtained on distorted datasets approach the performance obtained by recognition of undistorted signals of a similar nature. Compared to the baseline GMM-HMM, the WER value is reduced by 26.6% absolutely for highly compressed signals and by 29% for excessively denoised signals. This corresponds to a very large relative decrease in WER, namely, by 66% and 52%, respectively. These values of relative decrease in WER are more than two times higher than the typical values yielded by using DNN-HMM architecture on various speech data not affected by this type of nonlinear distortion (see [18] for example). In this case, we discovered no significant benefits of preprocessing by HEQ prior the recognition by DNN-HMM system.

The future work should focus on: 1) More detailed analysis of observed DNN robustness, e.g. through accuracy of phoneme recognition. 2) The investigation of influence of other preprocessing techniques (e.g. clipping compensation in [19]) on DNN-HMM recognition accuracy.

6. Acknowledgements

This work was supported by the Technology Agency of the Czech Republic (Project No. TA01011142).

7. References

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30-42, 2012.

[2] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14-22, 2012.

[3] A. Ghoshal, P. Swietojanski, S. Renals, "Multilingual training of deep neural networks." In Proc. *ICASSP 2013*, pp. 7319-7323, 2013.

[4] M. Seltzer, D. Yu, Y. Wang, "An investigation of deep neural networks for noise robust speech recognition", in Proc. *ICASSP 2013*, pp. 7398-7402, 2013.

[5] N. Parihar and J. Picone, "Aurora working group: DSR front-end LVCSR evaluation AU/384/02," *Tech. Rep.*, Inst. for Signal and Information Process, Mississippi State University.

[6] B. Li, Y. Tsao, K. Ch. Sim, "An investigation of spectral restoration algorithms for deep neural networks based noise robust speech recognition," in Proc. *Interspeech 2013*, August 2013.

[7] J. Malek, J. Silovsky, P. Cerva, Z. Koldovsky, J. Nouza and J. Zdansky, "Compensation of nonlinear distortions in speech for automatic recognition," accepted for publication in Proc. *TSP 2014*, Berlin, 2014.

[8] A. De la Torre, A. Peinado , J. C. Segura, J. L. P. Cordoba, C. Benitez, A. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355-366, 2005.

[9] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75-98, 1998.

[10] J. C. R. Licklider, I. Pollack, "Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech", *The Journal of the Acoustical Society of America*, 20(1), pp. 42-51, 2005.

[11] K. Dogancay, "Blind compensation of nonlinear distortion for bandlimited signals," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 52, no. 9, pp. 1872 - 1882, 2005.

[12] P. Pollak and M. Behunek, "Accuracy of mp3 speech recognition under real-world conditions. experimental study," in *SIGMAP 2011*, pp. 5-10, 2011.

[13] R. Kneser, H. Ney, "Improved backing-off for M-gram language modeling", in Proc. *ICASSP 1995*, pp. 181-184,1995.

[14] P. C. Woodland and D. Povey , "Minimum phone error and i-smoothing for improved discriminative training," in Proc. *ICASSP 2002*, 2002, pp. 105-108.

[15] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in Proc. *SciPy*, 2010.

[16] F. Seide, G. Li, X. Chen, D. Yu, "Feature engineering in Context-Dependent Deep Neural Networks for conversational speech transcription", *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pp. 42-51, 2011.

[17] Y. Obuchi, R. M. Stern. "Normalization of time-derivative parameters using histogram equalization." in Proc. *Interspeech 2003*, 2003.

[18] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition", *Signal Processing Magazine*, vol. 29, no.6, pp. 82-97, 2012.

[19] E. James, N. A. Patrick, "Detection of clipping in coded speech signals," in Proc. *EUSIPCO 2013*, pp.1-5, 2013.