Interspeech 2018
2-6 September 2018, Hyderabad

Long Distance Voice Channel Diagnosis Using Deep Neural Networks

Zhen Qin, Tom Ko, Guangjian Tian

Huawei Noah's Ark Research Lab, Hong Kong, China
{qinzhen8, tian.guangjian}@huawei.com, tomkocse@gmail.com

Abstract

In long distance telephone networks, it is time-consuming to detect and locate the problematic devices. Although hints could be given from the types of distortion in the test calls, it is tedious to manually classify the distortion types from a large number of calls. In this paper, we present our work on using a deep neural network-based classifier to automatically detect and identify the type of distortion which often occurs in long distance calls. We verified our approach with data from real telecommunication networks and the results showed that our approach can achieve an average recall rate of 71% in classification. We believe our method can lead to a huge reduction of manpower and time in long distance voice channel troubleshooting.

Index Terms: acoustic event detection, distortion detector, DNN, network diagnosis

1. Introduction

In telecommunication, it is important for the service providers to maintain a reliable voice channel between users. Regular testings and troubleshooting of a long distance voice channel are expensive and time-consuming. A long distance voice channel is usually composed of a number of network devices. Malfunction in any of the devices may result in voice quality degradation. Since different types of devices, when they are not working well, might induce different types of distortion to the transmitted signal, knowing the distortion type in the calls helps to locate the possible problematic devices.

However, identifying the type of distortion is not trivial. As the fault from a device often occurs in a random manner, normally a large number of test calls have to be made to capture the possible distortion. Though the test calls could be generated automatically, nowadays, these calls still have to be listened manually to identify the type of distortion.

Speech Quality Assessment (SQA) methods, e.g. the Perceptual Evaluation of Speech Quality (PESQ) [1] and Perceptual Objective Listening Quality Assessment (POLQA) [2], have been proposed to solve similar speech quality problems. These methods measure the speech quality in the call. When the score falls below a threshold, the call is classified as one with quality problem. However, these methods are not widely adopted in long distance voice channel diagnosis because of two reasons. First, the accuracy of these methods in distortion detection is not high. Second, these methods cannot classify the type of distortion.

Recently, the fast development of deep neural networks (DNN) facilitates a lot of acoustic event detection tasks [3, 4, 5, 6]. In this paper, we present our work on using a deep neural network-based classifier to automatically detect and identify the type of distortion in long distance calls. We verified our approach with data from real telecommunication networks and the results showed that our approach can achieve an average recall rate of 71% in classification. As we are not aware of any previous work that uses distortion classification method in network diagnosis, the aims of this paper are to investigate our idea and to establish a baseline.

The paper is organized as follows: Section 2 describes the types of distortion and their relation with various network devices. Section 3 describes our proposed method. Section 4 describes the experimental setup. Section 5 presents the results and finally the conclusions are presented in Section 6.

2. Types of distortion in long distance calls

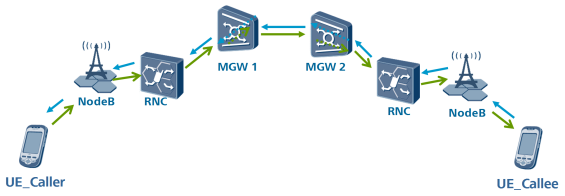

Figure 1: *A simplified view of a long distance Voice channel.*

Signal distortion is a common problem in long distance voice streaming which can negatively affect voice quality. This section describes the types of distortion that often occur in long distance calls and how they are related to the malfunction of network devices (Fig. 1). There are five major types of distortion that can trigger a check of network devices.

- **Low-loudness**: The main cause of this problem is the failure of gain in speech codecs. Thus, engineers need to check the devices which are related to speech codecs, such as the media gateway (MGW) shown in Fig. 1.
- **Noise**: When there is a parameter problem in any of the radio network controllers (RNC) or a codec problem in any of the MGWs, noises, especially metallic sounds and liquid sounds, will be induced to the transmitted speech. Thus, if these sounds are detected, engineers need to check the parameters of RNCs and the codecs of MGWs.
- **Discontinuity**: The main cause of discontinuity is usually from the air interfaces. Thus, engineers need to check the wireless devices, such as the base station NodeB shown in Fig. 1.
- **Low-quality**: Any content loss in the speech carried by the signal is defined as low-quality. This is most likely caused by the gateway configuration or the wireless coverage problem. Therefore, MGW's configuration and NodeB's coverage capability are checked.
- **Mute**: This could be induced by any devices on the channel. If mute is detected, engineers have to run further loopback tests to locate the failure devices.

10.21437/Interspeech.2018-1428

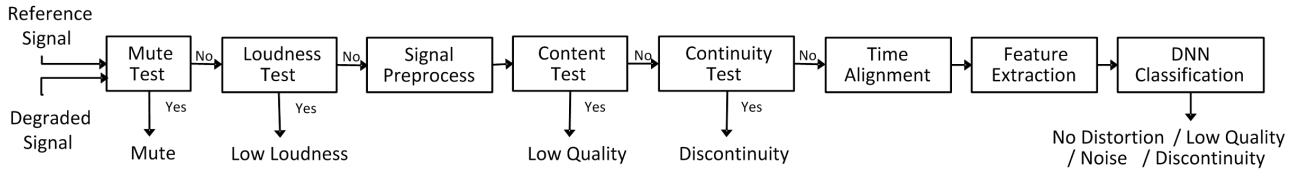

Figure 2: *Flow diagram of the proposed method*

3. Methodology

Fig. 2 summarizes our proposed method. It is composed of several signal processing blocks and a DNN classifier. In this paper, reference signals refer to speech signals at the caller side and degraded signals refer to speech signals at the receiver side. A degraded signal and its corresponding reference signal are used as input to our method. In real world applications, the test calls are made by the system engineers so they can always have the reference and degraded signals. As the main aim of our system is to detect the distortion induced by the channel so as to locate the problematic devices, the speech content of the signals is not important[1].

Our system can be viewed as a six-class classifier for the five distortion types plus no distortion. Regarding the five distortion types, mute and low-loudness are detected only by the signal processing blocks. Though there are processing blocks for detecting low-quality and discontinuity, they are not robust enough and the DNN classifier is needed to further classify the false negative samples. Nevertheless, these blocks are still needed to simplify the time alignment algorithm by reducing the complexity of the signals. Thus, the DNN and the signal processing blocks are both necessary in our system.

3.1. Mute & Loudness Tests

Given a pair of reference signal and degraded signal, a voice activity detection (VAD) [7] is applied to detect the mute. If a degraded signal does not contain any voice activity but its reference signal does, we regard this as a mute signal. After that, if the energy of a degraded signal is lower than its referenced signal by a factor of Θ_l, we regard this as a low-loudness signal.

3.2. Signal Preprocessing

If the degraded signal is not a low-loudness signal, it means the gain-effects between the signals satisfy the demand of human auditory. In order to normalize the gain induced by different telecommunication systems, the signals are aligned into the same power level and filtered by a band-pass filter. This emphasizes the information related to perception and could improve the performance [8].

3.3. Content & Continuity Tests

Here, a VAD algorithm is further used for splitting the signals into segments with flexible length. In our setup, every segment should contain at least 300ms of continuous voice activity and no more than 200ms of continuous silence[2]. If a silence with length more than 200ms exists in a segment, it will be split.

If the number of segments in a degraded signal is different from its reference signal, or the duration of any degraded segment is shorter than its corresponding reference segment by a factor of Θ_c, we regard this as a content loss.

After that, if the duration of silence within a degraded signal is much longer than the silence in its reference, we regard this as a discontinuity.

These blocks are important as they filter out the samples that are too hard for the time alignment block to fix.

3.4. Time Alignment

Here, the degraded signals are aligned with their reference signals to eliminate the delay induced by the network. This part is important as the feature extraction for DNN operates on the difference within the signal pair. As the time delay exists not only within a segment but also among segments, we divided time alignment into two parts: segment alignment and signal alignment.

For segment alignment, a degraded segment X needs to be further split into subsegments $\mathbf{x}$ at first. Since the envelope of a signal could be partitioned by troughs, we calculate the mean amplitude for each frame in 20ms and locate troughs whose mean amplitude is less than a threshold Θ_t. Then, the time delays between degraded subsegments $\mathbf{x}$ and their reference segment y are estimated. The cross-correlation $corr(x_i, y)$ of the most similar position between two signals is computed as follows:

$$corr(x_i, y) = \underset{n}{argmax}((x_i * y)(n)), 1 \leqslant n \leqslant N \quad (1)$$

$$(x_i * y)(n) = \frac{1}{M} \sum_{m=1}^{M} x_i(m+n)y(m) \quad (2)$$

where x_i is the i-th degraded subsegment in $\mathbf{x}$, y is the reference segment of $\mathbf{x}$, n is the offset of cross-correlation, N, M are the total numbers of sampling points within x_i and y respectively. Accordingly, the time delay τ_i between subsegment x_i and segment y is computed as follows:

$$\tau_i = corr(x_i, y) - p_{x_i} \quad (3)$$

where p_{x_i} is the starting position of x_i in degraded segment X. As the degraded signal has already passed the content and continuity tests, we assumed that the segments between a pair of signals are one-to-one, and the time delays within segments are small. Thus, if $\tau_i > \Theta_d$, τ_i will be abandoned. Based on the saved delays, the τ_i length of troughs will be removed before x_i in X if $\tau_i > 0$. Otherwise, the τ_i length of silence will be inserted before x_i in X.

Finally, for signal alignment, original degraded segments will be replaced by aligned degraded segments in the following ways:

[1] In our work, the test calls are made from prerecorded speech.

[2] We follow the setup in [8] to pick the length of silence.

- Insert or remove the silence at the head of a degraded signal till $p_{x_1'} = p_{y_1}$.
- If $p_{x_{i+1}} < p_{y_{i+1}}$ after replacing X_i with aligned $X_i^{'}$, insert the silence at the tail of $X_i^{'}$ till $p_{x_{i+1}} = p_{y_{i+1}}$.
- If $p_{x_{i+1}} \geqslant p_{y_{i+1}}$ after replacing X_i with aligned $X_i^{'}$, remove the silence at the tail of $X_i^{'}$ till $p_{x_{i+1}} = p_{y_{i+1}}$.
- Insert or remove the silence at the tail of a degraded signal till the aligned degraded signal has the same length with its reference.

where $p_{x_i'}$ and p_{y_i} are the starting positions of aligned $X_i^{'}$ and reference segment y_i respectively.

Table 1: *Low level descriptors (LLDs). The numbers between brackets are dimensions of the extracted feature vectors.*

MFCC(39), LPC(30), Loudness(24), Energy(1), Envelope(20), Envelope Shape Statistics(4)

Table 2: *Statistical functionals*

max, min, range, mean, maxmeandist, minmeandist, absmean, nzamean, nzabsmean, variance, stddev, rms, skewness, kurtosis, quartiles(3), iqr(3)

3.5. Feature Extraction

After the degraded signals are aligned with their references, feature vectors are extracted from each pair of signals and used as inputs to the neural network. Following the way in [9], We use the YAAFE toolbox [10] to extract 118 acoustic low-level descriptors (LLDs)[3] for the degraded and reference signals, as shown in Table 1. These LLDs [11] cover the spectral, cepstral, prosodic and voice quality information, which are relevant to our distortion detection. 20 functionals, as shown in Table 2, are computed by openSMILE [12] on the difference between degraded and reference LLDs, resulting in 2360 dimensional feature vectors. These feature vectors represent the degradation of the degraded signals from their references.

3.6. DNN Classifier

As parts of the distortion type are identified by the signal processing blocks, the DNN classifier only need to classify the samples into four classes: no distortion, noise, low-quality and discontinuity. Our DNN has 3 hidden layers of ReLU (rectified linear unit) nonlinearity [13] with input and output dimensions of 2360 and 4 respectively. It is trained by using cross-entropy [14] as the cost function.

4. Experimental Setup

We evaluate our method on both real data and simulated data but only simulated data are used to tune the system parameters and train the DNN model. In our task, real data means the recordings are distorted by real long distance channels and the simulated data[4] are generated by speech augmentation. The real set contains 101 samples, altogether 17.5 minutes; the simulated set contains 2,000 samples, altogether 5 hours. The simulated set is further divided into a training set of 1,500 samples and a test set of 500 samples. All samples are ranging from 7 to 20 seconds. Each sample contains a pair of degraded signal and reference signal. Each data set contains samples from the six classes: mute, low-loudness, discontinuity, noise, low-quality and no distortion.

For the system parameters, the threshold Θ_l for loudness test is 5 dBFS, threshold Θ_c for content test is 0.9, threshold Θ_t for trough location is 200, and threshold Θ_d for delay abandon is 0.125s. The numbers of units in the three DNN hidden layers are 1000, 3000 and 1000 respectively.

To analysis and estimate the performance of our experiments, the general measurements used in this paper are recall, precision and F1-score [16]. We also provide the confusion matrix for distortion classification.

5. Results

Table 3: *Performance of distortion detection.*

Test set	Recall	Precision	F1 Score
Simulated set	92.53%	96.65%	94.55%
Real set	94.02%	78.75%	85.71%

We first evaluate our method in distortion detection using simulated test set and real test set (Table 3). In this experiment, a positive means that a distortion is detected and a negative means that no distortion is found. We then evaluate the performance of our system on a six-class classification: no distortion (ND), mute (MU), noise (NOI), low-loudness (LL), discontinuity (DCT) and low quality (LQ).

Table 4: *Results in classification using simulated test set*

	Recall	Precision	F1 Score
ND	92.74%	84.56%	88.46%
LQ	76.92%	43.48%	55.56%
NOI	70.71%	87.50%	78.21%
DCT	62.79%	65.85%	64.29%
LL	92.59%	95.24%	93.90%
MU	100.0%	88.89%	94.12%
Average	**82.63%**	**77.59%**	**77.21%**

As shown in Table 4 and Table 5, our proposed method obtained an average F1 score of 77.21% on the simulated test set and the average recall rate is 82.63%. On the real test set, the average F1 score of our proposed method is 73.56% and the average recall rate is 71.92%. These numbers are different from those in Table 3 because they are computed from different aspects (detection vs. classification).

The confusion matrix is presented in Table 6. Among the six classes, discontinuity has the lowest recall rate (62.79%). This could be attributed to the existence of very long silences

[3]Since the envelope and envelope shape statistics are long-term descriptors, we extract these with 500 ms frame length and 10 ms frame shift, while all the other LLDs are extracted with 20 ms frame length and 10 ms frame shift.

[4]The noisy samples in the simulated set are generated by a superposition of separate noise signals and clean speech recordings [15].

Table 5: *Results in classification using real test set*

	Recall	**Precision**	**F1 Score**
ND	42.86%	78.95%	55.56%
LQ	42.86%	75.00%	54.55%
NOI	95.83%	62.16%	75.41%
DCT	64.29%	42.86%	51.43%
LL	85.71%	100.0%	92.31%
MU	100.0%	88.89%	94.12%
Average	**71.92%**	**74.64%**	**73.56%**

Table 6: *Confusion matrix of classification on simulated data*

Actual	Classified					
(%)	ND	LQ	NOI	DCT	LL	MU
ND	**92.74**	0.00	2.42	0.81	4.03	0.00
LQ	7.69	**76.92**	7.69	7.69	0.00	0.00
NOI	16.16	1.01	**70.71**	12.12	0.00	0.00
DCT	0.00	27.91	9.30	**62.79**	0.00	0.00
LL	3.70	0.00	1.85	0.00	**92.59**	1.85
MU	0.00	0.00	0.00	0.00	0.00	**100.0**

Table 7: *Recall rate of individual modules on simulated test set.*

	Continuity test module	**DNN**
DCT	53.85%	24.33%
	Content test module	**DNN**
LQ	68.19%	31.46%

in the degraded segments. Thus, much of the discontinuous signals (27.91%) are wrongly classified as low-quality.

Fig. 2 shows that two type of distortions, low quality and discontinuity, are handled in multiple modules. Table 7 reports the performance of individual modules on the simulated test set. In detecting discontinuity, around 54% of the target samples are detected and filtered out by the continuity test module and the remaining target samples are further processed by the DNN. Then the DNN can detect about 24% of the target samples from these samples of which the test module are failed to filter out. The recall rate of the DNN is lower than the test module because the DNN is handling more difficult samples.

6. Conclusions and Future work

In this paper, we evaluate our idea of using a DNN-based classifier to automatically detect and identify the type of distortion which often occurs in long distance calls. As knowing the distortion type gives an obvious hint to locate the possible problematic devices, we believe our method can lead to a huge reduction of manpower and time in long distance voice channel troubleshooting. We are investigating other neural network architectures for further improvement in performance.

7. References

[1] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of ICASSP*, vol. 2. IEEE, 2001, pp. 749–752.

[2] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part itemporal alignment," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[3] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015, pp. 1–7.

[4] O. Gencoglu, T. Virtanen, and H. Huttunen, "Recognition of acoustic events using deep neural networks," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*. IEEE, 2014, pp. 506–510.

[5] M. Espi, M. Fujimoto, Y. Kubo, and T. Nakatani, "Spectrogram patch based acoustic event detection and classification in speech overlapping conditions," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*. IEEE, 2014, pp. 117–121.

[6] Z. Kons and O. Toledo-Ronen, "Audio event classification using deep neural networks." in *INTERSPEECH*, 2013, pp. 1482–1486.

[7] D. Ying, Y. Yan, J. Dang, and F. K. Soong, "Voice activity detection based on an unsupervised learning framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2624–2633, 2011.

[8] A. W. Rix, M. P. Hollier, A. P. Hekstra, and J. G. Beerends, "Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part i–time-delay compensation," *Journal of the Audio Engineering Society*, vol. 50, no. 10, pp. 755–764, 2002.

[9] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 73–80.

[10] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software." in *ISMIR*, 2010, pp. 441–446.

[11] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV+EC 2015 - the first affect recognition challenge bridging across audio, video, and physiological data," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC)*. ACM, 2015, pp. 3–8.

[12] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[13] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[14] K.Vesely, A.Ghoshal, L.Burget, and D.Povey, "Sequence-discriminative training of deep neural networks." in *INTERSPEECH*, 2013, pp. 2345–2349.

[15] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5220–5224.

[16] D. M. W. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, pp. 37–63, 2011.