INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

End-to-End Deep Learning Framework for Speech Paralinguistics Detection Based on Perception Aware Spectrum

Danwei Cai[12], *Zhidong Ni*[12], *Wenbo Liu*[1], *Weicheng Cai*[1], *Gang Li*[3], *Ming Li*[12]

[1]School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China
[2]SYSU-CMU Shunde International Joint Research Institute, Guangdong, China
[3]Jiangsu Jinling Science and Technology Group Limited, Jiangsu, China

liming46@mail.sysu.edu.cn

Abstract

In this paper, we propose an end-to-end deep learning framework to detect speech paralinguistics using perception aware spectrum as input. Existing studies show that speech under cold has distinct variations of energy distribution on low frequency components compared with the speech under 'healthy' condition. This motivates us to use perception aware spectrum as the input to an end-to-end learning framework with small scale dataset. In this work, we try both Constant Q Transform (CQT) spectrum and Gammatone spectrum in different end-to-end deep learning networks, where both spectrums are able to closely mimic the human speech perception and transform it into 2D images. Experimental results show the effectiveness of the proposed perception aware spectrum with end-to-end deep learning approach on Interspeech 2017 Computational Paralinguistics *Cold* sub-Challenge. The final fusion result of our proposed method is 8% better than that of the provided baseline in terms of UAR.

Index Terms: computational paralinguistics, speech under cold, deep learning, perception aware spectrum

1. Introduction

Speech paralinguistics study the non-verbal signals of speech including accent, emotion, modulation, fluency and other perceptible speech phenomena beyond the pure transcriptional content of spoken speech [1]. With the advent of computational paralinguistics, such phenomena can be analysed by machine learning methods. The Interspeech COMPUTATIONAL PARALINGUISTICS CHALLENGE (COMPARE) is an open Challenge in the field of Computational Paralinguistics since 2009. Interspeech 2017 ComParE Challenge addressed three new problems within the field of Computational Paralinguistics: *Addressee* sub-challenge, *Cold* sub-challenge and *Snoring* sub-challenge [2].

In this paper, we proposed an efficient deep learning architecture for *Cold* sub-challenge of the Interspeech 2017 Computational Paralinguistics ChallengE [2]. The task aims to differentiate the cold-affected speech from the 'normal' speech. The baseline of challenge includes three independent systems. The first two systems use traditional classification method (i.e. SVM) with COMPARE features representation [3] and bag-of-audio-words (BoAW) features representation [4], and achieve unweighted average recall (UAR) of 64.0 and 64.2 respectively. The third system employs end-to-end learning but only achieves UAR of 59.1. Similar to [5], this system uses a convolutional network to extract features from the raw audio and then a subsequent recurrent network (i.e. LSTM) performs the final classification [2].

During the past few years, deep learning has made significant progress. Deep learning methods outperform the traditional machine learning methods in variety of speech applications such as speech recognition [6], language recognition [7], text-dependent speaker verification [8], emotion recognition [5], anti-spoofing tasks. This motivates us to apply deep learning methods to computational paralinguistic tasks.

However, the end-to-end baseline system provided in [2] did not achieve better UAR than the other two baseline systems. One possible reason is that small scale dataset may not be able to drive the deep neural network to learn a good feature directly from waveform for classification, and hard to obtain a robust feature for classification. We thus look into the frequency representation (i.e spectrograms) to perform the end-to-end learning. Spectrograms is a widely used audio signal feature representation in deep learning, which contain more wealth of acoustic information.

Existing study shows that compared with speech in 'health' condition, the speech in cold has larger amplitude in low frequency components and lower amplitude in high frequency components. [9]. Also, from the viewpoint of a human auditory perceptual system, human ears are more sensitive to small changes in low frequencies [10]. This motivates us to use perception aware spectrograms (i.e. Gammatone spectrograms and Constant Q Transform spectrograms) as the input for end-to-end deep learning framework when performing computational paralinguistics tasks. Constant Q transform employs geometrically spaced frequency bins and ensures a constant Q factor across the entire spectrum. This results in a finer frequency resolution at low frequencies while provides a higher temporal resolution at high frequencies [11]. Gammatone spectrum employs Gammatone filters which are conceived as a simple fit to experimental observations of the mammalian cochlea, and have a repeated pole structure leading to an impulse response that is the product of a gamma envelope $g(t) = t^n e^{-t}$ and a sinusoid (tone) [12, 13].

To the best of our knowledge, deep learning framework with CQT spectrograms input has been successfully applied to piano music transcription [14], audio scene classification and domestic audio tagging [15]. But the performance of deep learning framework with Gammatone spectrograms input still

This research was funded in part by the National Natural Science Foundation of China (61401524), Natural Science Foundation of Guangdong Province (2014A030313123), Natural Science Foundation of Guangzhou City (201707010363), the Fundamental Research Funds for the Central Universities (15lgjc12), National Key Research and Development Program (2016YFC0103905) and IBM Faculty Award.

http://dx.doi.org/10.21437/Interspeech.2017-1445

remains to be investigated. In this work, we try different network architecture with the above two perception aware spectrum, and find that perception aware spectrum outperforms the conventional short-term Fourier Transform (STFT) spectrum in the paralinguistic speech tasks of cold-affected speech. We think that our proposed method is applicable to other computational paralinguistic speech tasks as well.

The remainder of this paper is organized as follows. In next section, we will describe the proposed methods and background on its major components. Section 3 presents the dataset and experimental results. A brief conclusion is given in section 4.

2. Methods

2.1. Perception aware spectrum

2.1.1. STFT spectrograms

Traditionally, discrete-time short-term Fourier transform is used to generate spectrograms of the time representation audio signals. Actually, the STFT is a filter bank. The Q factor defined as the ratio between the center frequency f_k and the frequency bandwidth Δf is a measure of the selectivity of each filter:

$$Q = \frac{f_k}{\Delta f} \tag{1}$$

In STFT, the Q factor increases with the frequencies since the bandwidth Δf related to the window function is identical for all filters. However, human's ears can easily perceive small changes of low frequencies, but for high frequencies only gross differences can be detected. Human perception system is known to approximate a constant Q factor between 500Hz and 20kHz [10]. As a result, STFT spectrum with varied Q may not be good enough for speech signals analysis but perception aware spectrum can provide more discriminant information for cold-affected speech detection and other computational paralinguistic tasks.

2.1.2. CQT spectrograms

The first perception aware spectrum we try in the end-to-end deep learning framework is constant Q transform spectrograms. It was introduced by Youngberg and Boll [16] in 1978 and refined by Brown [17] some years later in 1991. In contrast to the fixed time-frequency resolution of Fourier methods, CQT ensures a constant Q factor across the entire spectrum and thus gives a higher frequency resolution for low frequencies and a higher temporal resolution for high frequencies.

The CQT $X(k, n)$ of a discrete time signal $x(n)$ can be calculated by

$$X(k,n) = \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j) a_k^*(j-n+N_k/2) \tag{2}$$

where k is the index of the frequency bins, N_k is a variable window lengths and $a_k(n)$ are complex-valued waveforms, here also called time-frequency atoms, which are defined as

$$a_k(n) = \frac{1}{C} w(\frac{n}{N_k}) \exp\left[i\left(2\pi n \frac{f_k}{f_s} + \Phi_k\right)\right] \tag{3}$$

where f_k is the center frequency of the corresponding frequency bin, f_s is the sampling rate, $w(t)$ is a window function and Φ_k is a phase offset. C is a scaling factor given by

$$C = \sum_{l=-\lfloor N_k/2 \rfloor}^{\lfloor N_k/2 \rfloor} w\left(\frac{l+N_k/2}{N_k}\right) \tag{4}$$

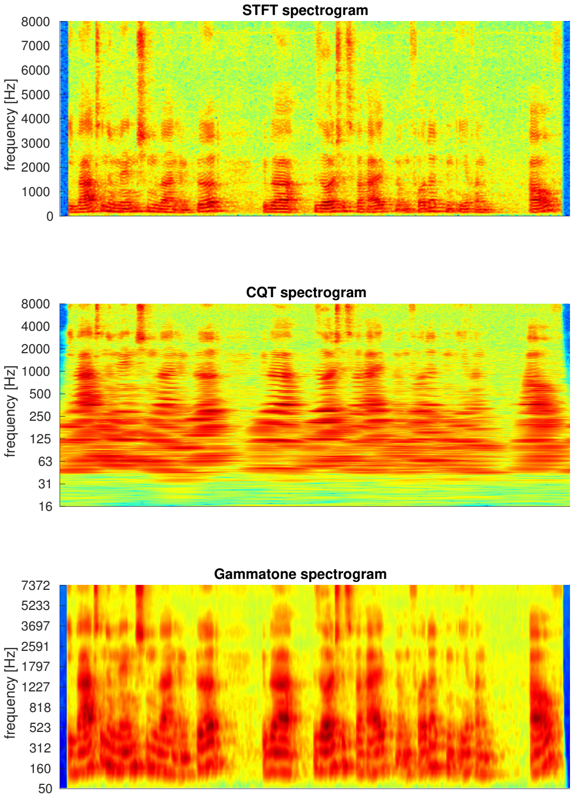

Figure 1: *Spectrograms of 'train_0250.wav' in URTIC dataset. Spectrograms computed with the short time Fourier Transform (top), with the constant Q transform (middle) and with Gammatone filters (bottom).*

Since a bin spacing corresponding to the equal temperament is desired, the center frequencies f_k obey

$$f_k = f_1 2^{\frac{k-1}{B}} \tag{5}$$

where f_1 is the center frequency of the lowest-frequency bin and B is a constant determines the time-frequency resolution trade-off. We then can write the Q factor as

$$Q = \frac{f_k}{f_{k+1} - f_k} = \left(2^{1/B} - 1\right)^{-1} \tag{6}$$

We can finally write the window lengths N_k which is inversely proportional to f_k to ensure a constant Q for all frequency bins as

$$N_k = \frac{f_s}{f_k} Q \tag{7}$$

2.1.3. Gammatone spectrograms

The second perception aware spectrum we try in the end-to-end deep learning framework is Gammatone spectrograms. Gammatone filters are a linear approximation to the filtering performed by the ear. To get a Gammatone spectrum, the audio signal is first analysed using a multi-channel Gammatone filterbank [18] and then the energy across each time frames is summed up [12].

Figure 1 shows STFT, CQT-derived spectrogram and Gammatone spectrogram for an arbitrarily selected speech signal from the dataset of the *Cold* sub-challenge. It is obvious that

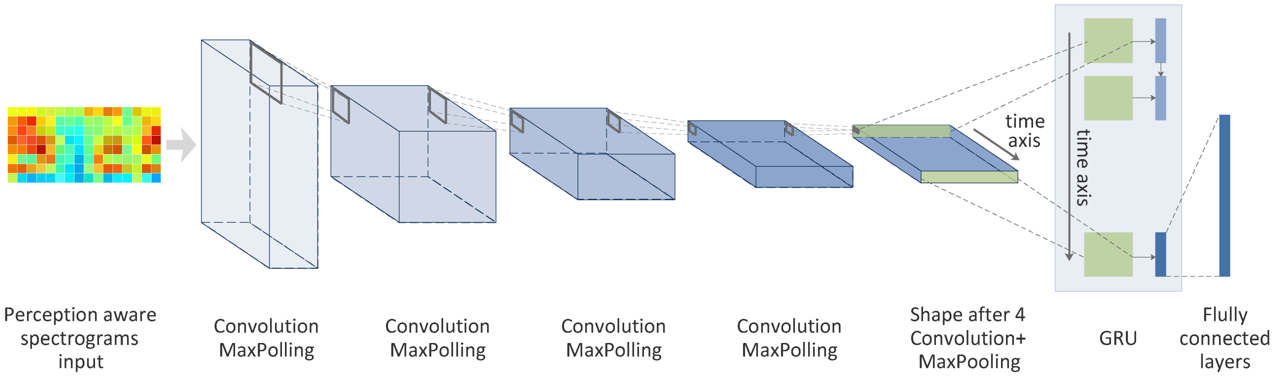

Figure 2: *The end-to-end network architecture with perception aware spectrograms input. The deep learning network consists of 4 convolution layers, 1 GRU layer and 1 fully connected layer.*

both CQT spectrum and Gammatone spectrum emphasize the low frequencies. The major difference between CQT spectrum and Gammatone spectrum is their low frequencies components. CQT spectrum gives a good frequency resolution but a bad time resolution, for it ensures a constant Q factor. Gammatone spectrum provides a smoother frequency resolution as human cochlea and a relatively good time resolution, for it apply Gammatone filters within each regular time bins. It's hard to say which kind of spectrogram will be better for the cold-affected speech detection tasks as well as other computational paralinguistic tasks. But one thing for sure is that perception aware spectrograms which reflect more closely the human perception system will provide more information in low frequencies and help the deep learning neural network to learn discriminate features for classification.

2.2. End-to-end deep learning framework

To perform end-to-end learning in cold-affected speech detection task, we combine convolutional neural network (CNN) and recurrent neural networks (RNN) to learn features automatically.

The general combinations scheme is as follows. First, a convolutional neural network acts as the feature extractor on the input perception aware spectrum input. Then, the CNN's output is feed into a recurrent neural network. The output of the CNN is a set of channels (i.e. feature maps). In our network, the 3-D tensor output of CNN is interpreted as a set of 2D-tensors along the time axis and each 2D-tensor contains the information from every channel. We employ a single gated recurrent unit (GRU) layer on 2D slices of that tensor and this enable the information from different channels mix inside the GRU. Finally, a fully connected layer with a softmax layer performs on the RNN's output to do classification. Figure 2 illustrates our end-to-end network architecture.

The CNN-LSTM deep learning framework has been successfully applied in the paralinguistic task of detecting spontaneous or natural emotions in a speech, except this work use a raw time representation input [5]. This framework with some residual connections ("shortcuts") from input to RNN and from CNN to fully connected layers has also been use in speech recognition [19].

2.3. CQCC and MFCC in GMM framework

To verify the effectiveness of the end-to-end deep learning network upon perception aware spectrum, we use CQCC and MFCC as the perception aware features to train classifier.

CQCC is based on constant Q transform which is already perception aware. The constant Q cepstral coefficients (CQCCs) of a discrete time signal with CQT $X(k)$ at a particular frame can then be extracted according to:

$$CQCC(p) = \sum_{l=1}^{L} \log |X(l)|^2 \cos \left[\frac{p\left(l - \frac{1}{2}\right)\pi}{L} \right] \tag{8}$$

where $p = 0, 1, \cdots, L - 1$ and l are the newly resampled frequency bins [11].

For MFCC, the Mel-cepstrum applies a frequency scale based on auditory critical bands before cepstral analysis [20]. The Mel-frequency cepstral coefficients (MFCCs) of a discrete time signal with DFT $X(k)$ at a particular frame can then be extracted according to:

$$MFCC(q) \sum_{m=1}^{M} \log \left[MF(m)\right] \cos \left[\frac{q\left(m - \frac{1}{2}\right)\pi}{M} \right] \tag{9}$$

where the $MF(m)$ is the Mel-frequency spectrum and is defined as

$$MF(m) = \sum_{k=1}^{K} |X(k)|^2 H_m(k) \tag{10}$$

where k is the DFT index and $H_m(k)$ is the triangular weighting-shaped function for the m-th Mel-scaled bandpass filter.

Two Gaussian mixture models (GMMs) is trained on one kind of perception aware features and used as a 2-class classifier in which the classes correspond to cold-affected and normal speech. The final score of a given test speech is computed as the log-likelihood ratio of the two GMMs.

3. Experiments

3.1. Dataset

We use the UPPER RESPIRATORY TRACT INFECTION CORPUS (URTIC) provided by the Institute of Safety Technology, University of Wuppertal, Germany. The corpus consists of 28652 instances with a duration between 3 and 10 seconds.

Table 1: *End-to-end learning network architecture. FC: fully connected layer. conv: convolutional layer.*

Network	Configuration
CNN+GRU+FC	conv1: 16 7×7 kernels, 1 stride conv2: 32 5×5 kernels, 1 stride conv3: 32 3×3 kernels, 1 stride conv4: 32 3×3 kernels, 1 stride pooling: 3×3 pool, 2×1 stride GRU: 500 hidden units FC: classification layer
CNN+FC	conv: same as above pooling: same as above FC1: 50 hidden units FC2: classification layer

9505 instances were selected for training, 9596 for the development set, and 9551 for testing.

The URTIC corpus is imbalanced: the number of cold-affected speech for training is 970 but the number of 'healthy' speech is 8535 [2]. However, a neural network trained on an imbalanced dataset may not be discriminative enough between classes [21]. To address this issue, we employ the simplest re-sampling technique by over-sampling the minority class with duplication when training end-eo-end deep learning networks.

3.2. Experimental results

We first use CQCC features to model the cold-affected speech and normal speech by employing two 512-components Gaussian mixture models and calculate the log-likelihood ratio upon these two GMMs for each test speech. We also use MFCC features with the same setup. The UAR with CQCC and MFCC features are 65.4% and 64.8% respectively, which is slightly better than the challenge organizer's SVM based results.

We then apply STFT spectrum, CQT spectrum and Gammatone spectrum to different end-to-end learning networks. Firstly, the training data is cut into a series of 3 seconds speech with an overlap of 2 seconds. We then extract different kinds of spectrograms which are 256×186 STFT spectrograms, 863×352 CQT spectrograms and 128×298 Gammatone spectrograms, the column number of these three spectrums are different due to the different frame shift parameters. All of which are used as input for the neural network. See table 1 for the details of the network architecture.

During neural network training phase, we use batch normalization to speed up. As the data are fed forward into a deep network, the parameters of the current layer adjust the input data and change the input data distribution for the next layer, which refers to as internal covariate shift. Batch normalization addresses the problem of internal covariate shift by normalizing layer inputs [22]. We also employ dropout to counter overfitting in training the neural network when labelled data is scarce [23].

Table 2 shows the experimental results of the baseline system and our proposed systems. It is observed that both CQT spectrum and Gammatone spectrum outperform the STFT spectrum in the case of UAR with the CNN+GRU+FC network setup. The best result of our end-to-end system (CQT spectrum with CNN+GRU+FC) is 15.7% better than the provided end-to-end network (raw waveforms with CNN+LATM). We use BOSARIS toolkit[24] to fuse the system results. The fusion results show that CQT and Gammatone spectrum are complementary to each other, and so does different neural network architectures. GMM system with CQCC or MFCC also helps to improve the system performance. The final fusion result of the URTIC development set is 71.4% and is 8% better than that of the provided baseline. The final fusion result of the test set, which is 66.71%, unfortunately suffers overfitting. We fuse it with the ComParE functional baseline (70.2%)[2] and get 71.2% UAR of the test set.

Table 2: *URTIC development set results for predicting the cold-affected speech.*

Algorithms	ID	Inputs	UAR
SVM [2]	1	ComParE functional	64.0%
	2	ComParE BoAW	64.2%
GMM	3	MFCC	64.8%
	4	CQCC	65.4%
CNN + LSTM [2]	5	Time representation	59.1%
CNN + FC	6	STFT spectrum	64.1%
	7	CQT spectrum	68.5%
	8	Gammatone spectrum	65.6%
CNN + GRU + FC	9	STFT spectrum	66.7%
	10	CQT spectrum	68.4%
	11	Gammatone spectrum	67.7%
Fusion	-	1+2+5 [2]	66.1 %
	-	6+7+8	68.7%
	-	9+10+11	69.9%
	-	7+8+10+11	70.8%
	-	6+7+8+9+10+11	70.6%
	-	3+4+6+7+8+9+10+11	**71.4**%

4. Conclusion

In this paper, we propose to use perception aware spectrum in end-to-end deep neural network to perform the computational paralinguistic task of detecting cold-affected speech. In the small scale datasets, perception aware spectrum such as CQT spectrum and Gammatone spectrum outperforms the raw time domain representation even the conventional STFT spectrum in end-to-end learning. We also investigate the performance of perception aware feature such as CQCC and MFCC when feeding it into GMMs which serve as a classification algorithm and verify the effectiveness of deep learning network with proper designed architecture and perception aware spectrum input. We have tried different spectrum input in different neural network architectures as well as the conventional classifier, fusing the results of these system brings a performance gain and shows that these features and methods are significant complementary to each other.

The computational paralinguistic task of detecting cold-affected speech still remains many problems to be investigated. For example, we have tried to use a phone decoder upon the given dataset and separately model three kinds of phone set which consist of vowel, nasal and other consonant with the split utterance. The experimental results show little discrimination between the three phone model mentioned above. This may due to the inaccurate phone decoder as well as the imbalanced phone set model training data. In the further work, we will try more accurate phone decoder and more proper modeling algorithms. Moreover, we will try to combine this idea with attention based neural network.

5. References

[1] B. Schuller, "The computational paralinguistics challenge [social sciences]," *IEEE Signal Processing Magazine*, vol. 29, pp. 97–101, 2012.

[2] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Proceedings of INTERSPEECH*, 2017.

[3] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings of INTERSPEECH*, 2013.

[4] M. Schmitt and B. W. Schuller, "openXBOW-introducing the passau open-source crossmodal bag-of-words toolkit," *preprint arXiv:1605.06778*, 2016.

[5] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proceedings of ICASSP*, 2016, pp. 5200–5204.

[6] 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," *Journal of IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[7] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, and P. J. Moreno, "Automatic language identification using long short-term memory recurrent neural networks." in *Proceedings of INTERSPEECH*, 2014, pp. 2155–2159.

[8] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proceedings of ICASSP*, 2016, pp. 5115–5119.

[9] Y. Shan and Q. Zhu, "Speaker identification under the changed sound environment," in *Proceedings of ICALIP*, 2014, pp. 362–366.

[10] B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[11] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proceedings of Speaker Odyssey Workshop, Bilbao, Spain*, vol. 25, 2016, pp. 249–252.

[12] D. Ellis, "Gammatone-like spectrograms," *web resource: www.ee.columbia.edu/dpwe/resources/matlab/gammatonegram*, 2009.

[13] Y. Shao, S. Srinivasan, and D. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *Proceedings of ICASSP*, vol. 4, 2007, pp. IV–277.

[14] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, pp. 927–939, 2016.

[15] T. Lidy and A. Schindler, "CQT-based convolutional neural networks for audio scene classification and domestic audio tagging," *DCASE 2016 Challenge*, 2016.

[16] J. Youngberg and S. Boll, "Constant-Q signal analysis and synthesis," in *Proceedings of ICASSP*, vol. 3, 1978, pp. 375–378.

[17] J. C. Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[18] M. Cooke, *Modelling Auditory Processing and Organization: Distinguished Dissertations in Computer Science Series*. Cambridge University Press Cambridge, 1993.

[19] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proceedings of ICASSP*, 2015, pp. 4580–4584.

[20] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *Proceedings of ICASSP*, vol. 28, 1980, pp. 357–366.

[21] E. DeRouin, J. Brown, H. Beck, L. Fausett, and M. Schneider, "Neural network training on unequally represented classes," *Intelligent engineering systems through artificial neural networks*, pp. 135–145, 1991.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of ICML*, 2015.

[23] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[24] N. Brmmer and E. D. Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new DCF," in *NIST SRE Analysis Workshop*, 2011.