*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# Automatic Estimation of Pathological Voice Quality based on Recurrent Neural Network using Amplitude and Phase Spectrogram

*Shunsuke Hidaka*[1,*], *Yogaku Lee*[2], *Kohei Wakamiya*[3], *Takashi Nakagawa*[2], *Tokihiko Kaburagi*[3,†]

[1]Graduate School of Design, Kyushu University, Japan
[2]Department of Otorhinolaryngology, Faculty of Medicine, Kyushu University, Japan
[3]Faculty of Design, Kyushu University, Japan

*hidaka.shunsuke.323@s.kyushu-u.ac.jp, †kabu@design.kyushu-u.ac.jp

## Abstract

Perceptual evaluation of voice quality is widely used in laryngological practice, but it lacks reproducibility caused by inter- and intra-rater variability. This problem can be solved by automatic estimation of voice quality using machine learning. In the previous studies, conventional acoustic features, such as jitter, have often been employed as inputs. However, many of them are vulnerable to severe hoarseness because they assume a quasi-periodicity of voice. This paper investigated non-parametric features derived from amplitude and phase spectrograms. We applied the instantaneous phase correction proposed by Yatabe *et al.* (2018) to extract features that could be interpreted as indicators of non-sinusoidality. Specifically, we compared log amplitude, temporal phase variation, temporal complex value variation, and mel-scale versions of them. A deep neural network with a bidirectional GRU was constructed for each item of GRBAS Scale, a hoarseness evaluation method. The dataset was composed of 2545 samples of sustained vowel /a/ with the GRBAS scores labeled by an otolaryngologist. The results showed that the Hz-mel conversion improved the performance in almost all the case. The best scores were obtained when using temporal phase variation along the mel scale for Grade, Rough, Breathy, and Strained, and when using log mel amplitude for Asthenic.

**Index Terms**: voice disorder, GRBAS Scale, short-time Fourier transform, phase correction, recurrent neural network

## 1. Introduction

Speech medicine in the field of otolaryngology deals with various voice disorders: laryngeal cancer, vocal cord paralysis, vocal cord polyps, and functional dysphonia. Hoarseness is not only the result of some voice disorder, but it can also cause a decrease in quality of life (QOL). Therefore, the evaluation of voice quality is an important diagnostic item. There are two methods for evaluating pathological voice quality: auditory-perceptual evaluation and acoustic analysis. In clinical practice, perceptual evaluation is used more often because voice quality is fundamentally perceptual in nature [1].

Acoustic analysis is an objective method of quantitative voice quality evaluation using acoustic features considered to be related to hoarseness. To date, many features have been devised, including jitter, shimmer, harmonic-to-noise ratio, spectral and cepstral features. Many of them assume quasi-periodicity of voice; hence they are vulnerable to non-stationary severe hoarse voice [2]. In addition, it is difficult to interpret them intuitively in relation to auditory impression. Therefore, acoustic analysis is an adjunct to auditory psychological evaluation. In contrast, auditory-perceptual evaluation is subjective; thus, this method inevitably lacks of reproducibility owing to inter- and intra-rater variation [3, 4].

GRBAS scale [5] is one of the most accepted auditory-perceptual evaluation methods. The scale consists of five measures: grade of hoarseness (G), rough (R), breathy (B), asthenic (A), and strained (S). For each item, clinicians provide a score of 0 (normal), 1 (slight), 2 (moderate) or 3 (severe). It has been suggested that items G, R and B are relatively reproducible when evaluated by a skilled person, whereas the reproducibility is lower for A and S [3, 4].

The problem of auditory psychological evaluation can be solved by automatic estimation based on machine learning. Computer-based estimation ensures reproducibility. Besides, if the automatic estimation system can evaluate voice quality at the same level as clinicians, then its validity can be considered sufficient. Many of the previous studies have used acoustic features derived from acoustic analysis [6, 7]. However, as mentioned earlier, many of the features are parametric and are not robust to severe hoarse voices. On the other hand, in the fields of acoustic scene classification and speech recognition, high-dimensional data such as spectrograms have been used as input features [8]. The handling of high-dimensional data has been made possible by deep learning. A similar approach could be extended to the evaluation of pathological voice quality.

The purpose of this study is to investigate the effectiveness of deep learning with spectrogram input for the evaluation of pathological voice quality. Although only the amplitude of spectrogram is often used [8], several studies have suggested that phase may contain useful information about hoarseness [9, 10]. Therefore, both amplitude and phase should be investigated. In this paper, we compared six spectrogram variations: amplitude, temporal variation of the phase or complex spectrogram, and a Hz-mel converted version of them. For the investigation, we constructed a deep neural network to output GRBAS scores using information derived from the spectrogram of sustained vowel /a/ as input.

## 2. Methods

### 2.1. Spectrogram

The short-time Fourier transform (STFT) of a signal is ordinarily called a complex spectrogram. In this paper, a (complex) spectrogram is defined as an *instantaneous phase corrected* STFT (iPC-STFT) of a signal proposed by Yatabe *et al* [11, 12, 13]. The instantaneous phase correction facilitates intuitive interpretation and technical application of the phase. This study considered the amplitude and phase of spectrograms defined above. Also, we investigated the application of mel-frequency scale to the phase, which has been done only for the amplitude thus far.

http://dx.doi.org/10.21437/Interspeech.2020-3228

### 2.1.1. *Instantaneous phase correction*

Instantaneous phase correction [11, 12, 13] is a method that focuses on the properties of the phase spectrogram of sinusoidal waves. So, let us first investigate its nature. The following explanation should be referred to [11, 12, 13].

The STFT of a signal $x(t)$ with respect to a fixed real window function $w(t)$ is defined as

$$\mathcal{F}^w x(m,n) = \sum_{l=an}^{an+L-1} x(l)w(l-an)e^{-2\pi i m l/L}, \tag{1}$$

where $i = \sqrt{-1}$, $L \in \mathbb{N}$ is the length of $w(t)$, $n \in \mathbb{N}$ and $m \in \mathbb{N}$ are the time and frequency indices respectively, and $a \in \mathbb{N}$ is the time-shifting step. For a complex sinusoid $x(l) = e^{2\pi i \xi l/L}$ where $\xi \in \mathbb{R}$, the following relationship holds when there exists a certain frequency index $m$ such that $\xi = m$:

$$\mathcal{F}^w x(\xi, n+1) = \mathcal{F}^w x(\xi, n) = \sum_{l=0}^{L-1} w(l). \tag{2}$$

However, if the condition is removed, $\xi$ cannot be assigned as a frequency index. For any frequency index $m$, the following relation holds:

$$\mathcal{F}^w x(m, n+1)e^{-2\pi i a\delta/L} = \mathcal{F}^w x(m,n), \tag{3}$$

where $\delta = \xi - m$ is the deviation; thus, the STFT's phase $\phi(m,n)$ (with appropriate unwrapping) has the following relation:

$$\phi(m, n+1) = \phi(m,n) + \frac{2\pi\delta}{L}. \tag{4}$$

The last term $2\pi\delta/L$ is caused by discretization. In fact, the deviation $\delta$ corresponds to the relative instantaneous frequency $\partial\phi/\partial n$. This indicates that, for the sinusoidal wave, the phase in the next frame is predictable from the phase and instantaneous frequency in the current frame.

Instantaneous phase correction aims to reduce the amount of mismatch between the frequencies of a sinusoid and frequency bins [12]. For this purpose, relative instantaneous frequency is used. The iPC-STFT of a signal $x(t)$ with respect to a fixed real window $w(t)$ was proposed in [11]:

$$\mathcal{F}^w_{\mathrm{iPC}} = A \odot \mathcal{F}^w, \tag{5}$$

where $A(m,n)$ is the instantaneous phase correction matrix defined by

$$A(m,n) = \prod_{\eta=0}^{n-1} \exp\left(-2\pi i \left.\frac{\partial\phi(m,n)}{\partial n}\right|_{n=\eta} /L\right) \tag{6}$$

with $A(m,0) = 1$ for all $m$, and $\odot$ is the Hadamard product (this notaion is according to [13]). As mentioned at the beginning of this subsection, we refer to an iPC-STFT as a (complex) spectrogram in particular.

For a signal composed of three sinusoidal waves with background noise, the STFT and the iPC-STFT are shown in Figure 1. In the frequency bins containing a frequency component of by the sinusoidal waves, the phase of STFT is rotated, while that of iPC-STFT is almost constant. In other words, concerning the iPC-STFT, the difference of the phase between the adjacent frames is almost zero in the sinusoid-dominated frequency bins, while it takes seemingly random values in the other bins. That is, instantaneous phase correction makes the existence of a sinusoidal component more distinctive; hence, instantaneous phase correction has been applied to some applications enhancing or distinguishing sinusoidal components such as speech enhancement [11] and harmonic/percussive source separation [13]. Incidentally, pathological voice is characterized by its non-stationarity, fluctuation, and high noisiness compared to the normal voice. Therefore, we could expect that the iPC-SPEC might include these characteristics as phase variations. This is why we adopted iPC-STFT instead of the usual STFT.

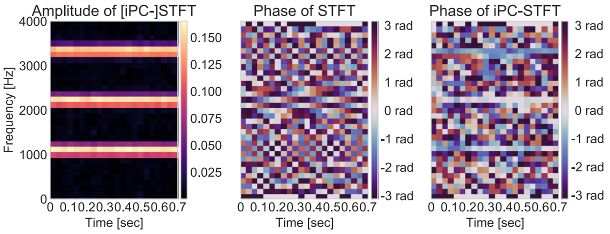

Figure 1: *Spectrograms of a signal composed of three sinusoidal waves with background noise. Note that the amplitude of the STFT and that of the iPC-STFT are the same. The color map for the phase spectrograms is cyclic to resolve the phase discontinuity. Due to the effect of instantaneous phase correction, the phase of the iPC-STFT is almost constant in the frequency bins affected by the sinusoidal waves.*

### 2.1.2. *Input features*

In this study, we considered the following spectrograms as the input features: log-amplitude spectrogram (LogAmp), absolute temporal difference of phase spectrogram (DiffPhase), and absolute temporal difference of complex spectrogram (DiffComplex). They are defined as follows:

$$\mathrm{LogAmp}^w x = 10\log_{10}|\mathcal{F}^w_{\mathrm{iPC}} x|^2, \tag{7}$$

$$\mathrm{DiffPhase}^w x = |D_t U_t \angle \mathcal{F}^w_{\mathrm{iPC}} x|, \tag{8}$$

$$\mathrm{DiffComplex}^w x = 10\log_{10}|D_t \mathcal{F}^w_{\mathrm{iPC}} x|^2, \tag{9}$$

where $w$ is a real window function, $D_t$ is the temporal difference $(D_t z)(m,n) = z(m,n) - z(m,n-1)$, $U_t$ is the temporal phase unwrapping, and $\angle$ is the angle. LogAmp is used in various applications and it can be intuitively interpreted. DiffPhase can be considered as a measure of non-sinusoidality. As already mentioned in the previous part, DiffPhase is almost zero in the sinusoid-dominated frequency bins, while it takes seemingly random values in other bins. We expected DiffPhase to contain information attributable to pathological voice characteristics such as non-stationarity. DiffComplex differs from the other two features in that it takes amplitude and phase into account simultaneously. DiffComplex appears in the computation of a feature called "phase corrected total variation (PCTV)" [11, 14]. The PCTV is defined as the $l^1$ norm of $D_t \mathcal{F}^w_{\mathrm{iPC}} x$. In fact, in order to improve the PCTV, instantaneous phase correction was introduced in [11]. DiffComplex is expected to reflect the presence of noise or fluctuating components. DiffPhase could be interpreted as the 'ratio' of non-sinusoidality to sinusoidality, while DiffComplex as the 'strength' of non-sinusoidality.

In addition to the above features, a Hz-mel conversion was also taken into consideration. The mel scale [15] is a perceptually proportional scale of pitch based on human hearing. The evaluation of pathological voice quality is exactly based on human hearing; thus, it is expected that the conversion would be an effective dimensionality reduction. Traditionally, the conversion has been applied only to amplitude spectrograms, but in this paper, it is also applied to the other two features. The mel versions are as follows: log mel amplitude spectrogram (MelLogAmp), mel absolute temporal difference of phase spectro-

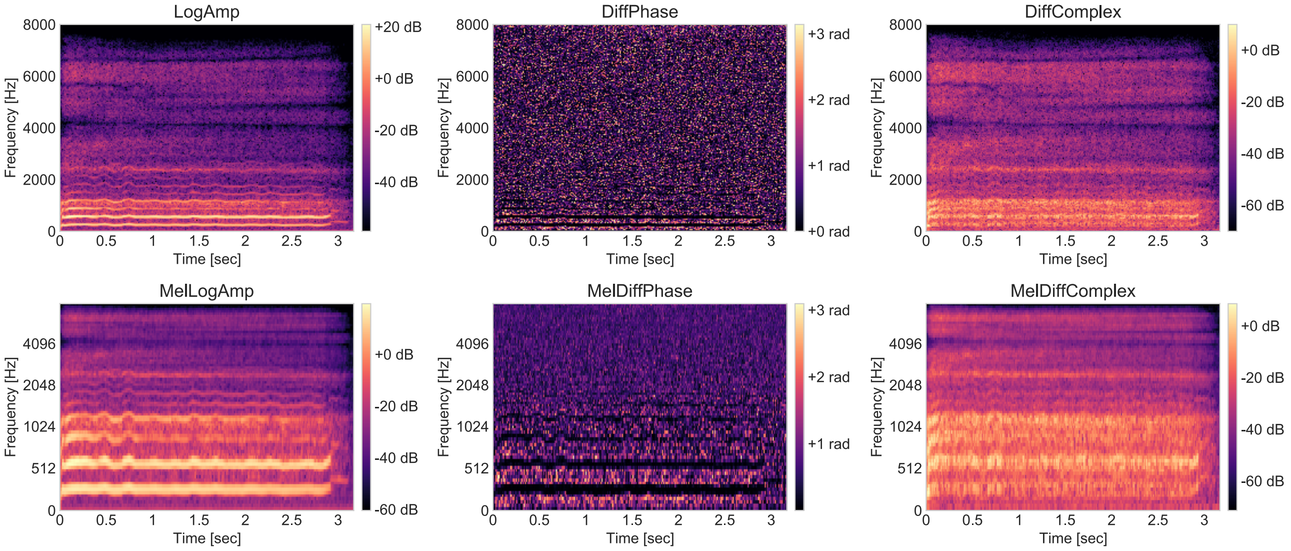

Figure 2: *The heatmap of the input features calculated from a sustained vowel /a/. The voice was rated as G3 R0 B3 A0 S1 by the doctor who gave all the scores of our dataset for the experiment.*

gram (MelDiffPhase), mel absolute temporal difference of complex spectrogram (MelDiffComplex). They are calculated by

$$\mathrm{MelLogAmp}^{w}x = 10\log_{10} M(|\mathcal{F}^{w}_{\mathrm{iPC}}x|^{2}), \tag{10}$$

$$\mathrm{MelDiffPhase}^{w}x = M|D_t U_t \angle \mathcal{F}^{w}_{\mathrm{iPC}}x|, \tag{11}$$

$$\mathrm{MelDiffComplex}^{w}x = 10\log_{10} M(|D_t \mathcal{F}^{w}_{\mathrm{iPC}}x|^{2}), \tag{12}$$

where $M$ is the Hz-mel conversion, and other symbols are the same as in Eq. (7)–(9).

Figure 2 shows the heatmap of the features calculated from a sustained vowel /a/. They were used as the input to a DNN described below.

### 2.2. Deep Neural Network

In this study, we constructed a DNN with a bidirectional GRU [16] and fully connected layers. Independent DNNs were constructed for each GRBAS item. They input the spectral sequences of input features and output a score of 0, 1, 2 or 3 as a solution to the classification problem.

Table 1 shows the block diagram of the DNN. The dynamic standardizer updates its running mean and variance each time a new input comes, and then uses them to standardize the input so that the standard deviation is 1 and the mean is 0. Note that the standardization is done independently for each frequency bin. The update rule of running statistics is as follows:

$$\hat{x}_{\mathrm{new}} = 0.9\hat{x}_{\mathrm{current}} + 0.1x_{\mathrm{input}}, \tag{13}$$

where $\hat{x}_{\mathrm{new}}$ is the new running statistic, $\hat{x}_{\mathrm{current}}$ is the current running one, and $x_{\mathrm{input}}$ is the observed value calculated from the new input. Dropout [17] and $l^2$ regularization are introduced to reduce overfitting, and gradient norm clipping [18] is applied to suppress gradient explosion [18]. We trained the DNN using the cross-entropy error as a loss function and Adam [19] as an optimizer.

## 3. Experimental Results

### 3.1. Database

We built a database consisting of 2545 samples (524 females and 859 males) of sustained vowel /a/. 2475 samples (504 females and 809 males) were pathological voice samples recorded at Kyushu University Hospital, and the remaining 70 samples (20 females and 50 males) were healthy voice samples recorded at the Ohashi Campus of Kyushu University. The causes of voice disorders included vocal cord paralysis, laryngeal cancer, vocal cord polyps, and spasmodic dysphonia. Although the recording environment was not consistent, each sample was recorded in a soundproof room with a microphone having a relatively flat frequency response. One expert otolaryngologist rated and labeled the GRBAS scores to each sample. The sound was played from headphones (MDR-CD900ST, SONY) through an audio interface (UA-25EX, cakewalk by Roland) connected to a PC. For the rating, the sampling frequency was set to 48 kHz (the original sampling frequency was either 48 kHz, 50 kHz, or 96 kHz).

Table 1: *The structure of the DNN. Data flows from top to bottom in forward propagation, and from bottom to top in backpropagation. The value in parentheses represents the dimension of output ($X$ is the total number of the frequency bins of one input feature, $N$ is the parameter for the hidden size). The output sequences of the forward and reverse layers of the bidirectional GRU are averaged over time and then combined as inputs to the subsequent layers. The structure of the "hidden layer" in the left table is as shown in the right. The last output dimension of the DNN is four in order to output the score of 0, 1, 2, or 3 as a solution to the classification problem.*

| **DNN** |
|---|
| *Input* ($X$) |
| Dynamic standardizer ($X$) |
| Fully connected ($N$) |
| Bidirectional GRU ($2N$) |
| Dropout ($2N$) |
| 0–2 hidden layers ($2N$) |
| Fully connected (4) |
| *Output* (4) |

| **A hidden layer** |
|---|
| *Input* ($2N$) |
| Fully connected ($2N$) |
| ReLU ($2N$) |
| Dropout ($2N$) |
| *Output* ($2N$) |

### 3.2. Experiments

From the database, we computed all the features described in Section 2.1.2. Before the feature extraction, the sampling frequency was downsampled to 16 kHz. For the framing of iPC-

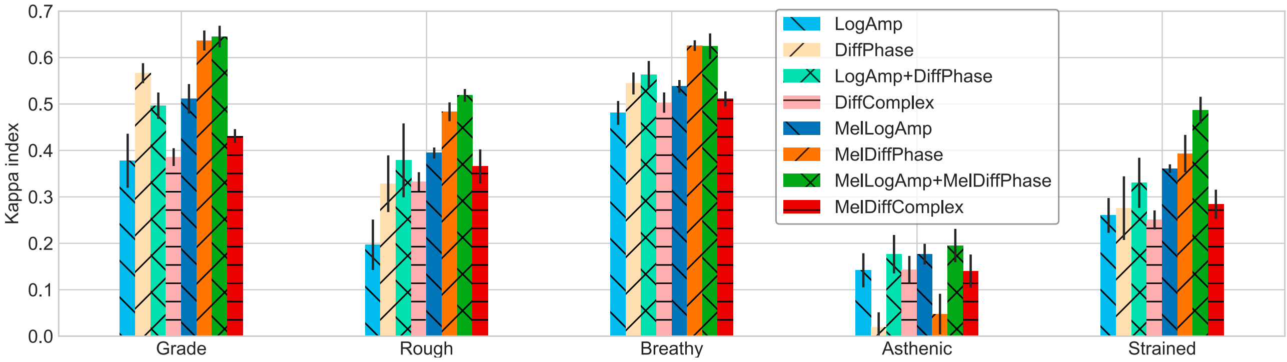

Figure 3: *The results from five-fold cross validation. The statistic on the vertical axis is linear weighted Cohen's kappa index $\kappa_L$. Error bars represent the standard deviation. The test dataset was evaluated using the model obtained through the training when the loss to the validation dataset was the minimum. The value for each GRBAS item and each feature was obtained using DNN parameter values with the highest $\kappa_L$.*

STFT, we used a 512-sample Hann window, with shifts of 128 samples. The mel filterbank consisted of 80 triangular filters. Each filter was normalized for the area ($l^1$ norm) to be one. As a definition of the mel scale, the implementation by Slaney [20] was used. The value of each spectrogram was normalized from zero to one. The total number of frequency bins was 257 for iPC-STFT and 80 for mel-scale features.

We constructed and tested the DNN models using PyTorch [21]. For some parameters of the DNN, We executed a grid search: the parameter $N$ for hidden size in Table 1 was either 128, 256, or 512, and the number of hidden layers was either two, three, or four. The parameters for the training were set as follows. The learning rate was 0.001, the weight of the $l^2$ regularization was 0.001, the gradient norm clipping threshold was 1.0, the dropout rate was 0.5, and the batch size was 128. The training was terminated when the minimum loss value was not updated for 15 epochs in succession.

For each GRBAS item, we compared the following eight features: LogAmp, DiffPhase, DiffComplex, MelLogAmp, MelDiffPhase, MelDiffComplex, the combined vector of LogAmp and DiffPhase (LogAmp+DiffPhase), the combined vector of MelLogAmp and MelDiffPhase (MelLogAmp+MelDiffPhase). Five-fold cross validation was used to evaluate the performance. We used 300 samples for the test and the rest samples were for training and validation (the speakers for the test and those for the training and validation were completely separated). Table 2 shows the distribution of the dataset.

Table 2: *The dataset distribution.*

| Division | Score | G | R | B | A | S |
|---|---|---|---|---|---|---|
| Test | 0 | 30 | 89 | 92 | 160 | 97 |
| | 1 | 90 | 110 | 102 | 97 | 101 |
| | 2 | 90 | 63 | 67 | 37 | 65 |
| | 3 | 90 | 38 | 39 | 6 | 37 |
| Training/ Validation | 0 | 198 | 634 | 698 | 1614 | 827 |
| | 1 | 1059 | 1071 | 867 | 544 | 1018 |
| | 2 | 614 | 386 | 431 | 81 | 293 |
| | 3 | 374 | 154 | 249 | 6 | 107 |

### 3.3. Results and Discussion

Figure 3 shows the experimental results. We evaluated the performance with linear weighted Cohen's kappa index $\kappa_L$ [22], which takes into account the possibility of the agreement occurring by chance. The $\kappa_L$ value indicates the strength of agreement, with a maximum of 1. For all GRBAS items, Hz-mel conversion improved the performance; except for DiffComplex and MelDiffComplex for item A. The comparison of features with Hz-mel conversion was as follows. The highest values for items G, R, A, and S were obtained by MelLogAmp+MelDiffPhase; that for item B by MelDiffPhase. The lowest values for items G, R, B, and S were obtained by MelDiffComplex; that for item A by MelDiffPhase. In the Landis and Koch's benchmark [23], the best scores in Figure 3 are ranked as follows: substantial for G and B; moderate for R and S; slight for A. Also, when comparing MelLogAmp and MelDiffPhase, MelDiffPhase outperformed MelLogAmp for items G, R, B, and S, while MelLogAmp did MelDiffPhase for item A.

The Hz-mel conversion, which is a linear map to a lower dimension, could be represented by fully connected layers in the DNN. In other words, the 'raw' spectrogram without the Hz-mel conversion contains abundant information. Nevertheless, in the present study, we find that the Hz-mel conversion is effective in almost all cases. The results suggest that Hz-mel conversion is a dimensionality reduction that can retain important information for auditory impressions. Also, the size of our dataset might have been too small for the dimension of raw spectrograms.

In this paper, we showed that valuable information can be extracted from the phase using the instantaneous phase correction. Besides, for items G, R, B, and S, the phase information was more effective than the amplitude information. This suggests that the phase variation could contain information specific to hoarseness, such as noise, non-stationarity, and fluctuation. Only for item A, the amplitude information had more contribution. This implies that asthenic can be associated with amplitude. The complex value variation was less effective. We suspect that one of the reasons for this might be the ambiguity of the harmonic structure in the complex value variation, as seen in Figure 2.

## 4. Conclusions

In this paper, we investigated the automatic estimation of pathological voice quality using the DNN, which inputted the information derived from amplitude and phase spectrograms. The temporal variation of phase with the instantaneous phase correction was shown to be effective in the evaluation of pathological voice. The experimental results suggest the complementarity of amplitude and phase spectrograms and the importance of considering both together. Future work should exploit a convolutional neural network, that can utilize the amplitude and phase in the time-frequency plane.

# 5. References

[1] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research," *Journal of Speech and Hearing Research*, vol. 36, no. 1, pp. 21–40, 1993.

[2] P. Carding, J. A. Wilson, K. MacKenzie, and I. J. Deary, "Measuring voice outcomes: State of the science review," *Journal of Laryngology and Otology*, vol. 123, no. 8, pp. 823–829, 2009.

[3] T. L. Eadie and C. R. Baylor, "The Effect of Perceptual Training on Inexperienced Listeners' Judgments of Dysphonic Voice," *Journal of Voice*, vol. 20, no. 4, pp. 527–544, 2006.

[4] M. S. De Bodt, F. L. Wuyts, P. H. Van De Heyning, and C. Croux, "Test-retest study of the GRBAS scale: Influence of experience and professional background on perceptual rating of voice quality," *Journal of Voice*, vol. 11, no. 1, pp. 74–80, 1997.

[5] M. Hirano, "Psycho-acoustic evaluation of voice," *Clinical examination of voice : disorders of human communication*. New York: Springer-Verlag, pp. 81–84, 1981.

[6] Z. Wang, P. Yu, N. Yan, L. Wang, and M. L. Ng, "Automatic Assessment of Pathological Voice Quality Using Multidimensional Acoustic Analysis Based on the GRBAS Scale," *Journal of Signal Processing Systems*, vol. 82, no. 2, pp. 241–251, 2016.

[7] J. A. Gómez-García, L. Moro-Velázquez, J. Mendes-Laureano, G. Castellanos-Dominguez, and J. I. Godino-Llorente, "Emulating the perceptual capabilities of a human evaluator to map the GRB scale for the assessment of voice disorders," *Engineering Applications of Artificial Intelligence*, vol. 82, pp. 236–251, 2019.

[8] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification: an overview of dcase 2017 challenge entries," *Proc. IWAENC*, pp. 411–415, 2018.

[9] M. Koutsogiannaki, O. Simantiraki, G. Degottex, and Y. Stylianou, "The Importance of Phase on Voice Quality Assessment," *Proc. INTERSPEECH*, 2014.

[10] T. Drugman, T. Dubuisson, and T. Dutoit, "Phase-based information for voice pathology detection," *Proc. ICASSP*, pp. 4612–4615, 2011.

[11] K. Yatabe and Y. Oikawa, "Phase Corrected Total Variation for Audio Signals," *Proc. ICASSP*, pp. 656–660, 2018.

[12] K. Yatabe, Y. Masuyama, T. Kusano, and Y. Oikawa, "Representation of complex spectrogram via phase conversion," *Acoustical Science and Technology*, vol. 40, no. 3, pp. 170–177, 2019.

[13] Y. Masuyama, K. Yatabe, and Y. Oikawa, "Phase-aware Harmonic/percussive Source Separation via Convex Optimization," *Proc. ICASSP*, pp. 985–989, 2019.

[14] İ. Bayram and M. E. Kamasak, "A Simple Prior for Audio Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 6, pp. 1190–1200, 2013.

[15] S. S. Stevens, J. Volkmann, and E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[16] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *Proc. EMNLP*, pp. 1724–1734, 2014.

[17] N. Srivastava, G. 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, no. 1, pp. 1929–1958, 2014.

[18] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," *Proc. ICML*, pp. 1310–1318, 2013.

[19] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, pp. 1–15, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[20] M. Slaney, "Auditory Toolbox: A MATLAB Toolbox for Auditory Modeling Work Version 2," Technical Report 1998-010, Interval Res. Corp., 1998.

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *Advances in Neural Information Processing Systems*, pp. 8024–8035, 2019.

[22] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968.

[23] R. J. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.