Dysarthric Speech Recognition, Detection and Classification using Raw Phase and Magnitude Spectra

Yue, Zhengjun; Loweimi, Erfan; Cvetkovic, Zoran

DOI
10.21437/Interspeech.2023-222

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH

INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Dysarthric Speech Recognition, Detection and Classification using Raw Phase and Magnitude Spectra

Zhengjun Yue[1,2,†], *Erfan Loweimi*[1,3,†], *Zoran Cvetkovic*[1]

[1] Department of Engineering, King's College London, UK
[2] Multimedia Computing Group, Delft University of Technology, the Netherlands
[3] Speech Group, Machine Intelligence Laboratory, University of Cambridge
{zhengjun.yue,erfan.loweimi,zoran.cvetkovic}@kcl.ac.uk

Abstract

In this paper, we explore the effectiveness of deploying the raw phase and magnitude spectra for dysarthric speech recognition, detection and classification. In particular, we scrutinise the usefulness of various raw phase-based representations along with their combinations with the raw magnitude spectrum and filterbank features. We employed single and multi-stream architectures consisting of a cascade of convolutional, recurrent and fully-connected layers for acoustic modelling. Furthermore, we investigate various configurations and fusion schemes as well as their training dynamics. In addition, the accuracies of the raw phase and magnitude based systems in the detection and classification tasks are studied and discussed. We report the performance on the UASpeech and TORGO dysarthric speech databases and for different severity levels. Our best system achieved WERs of 31.2% and 9.1% for dysarthric and typical speech on TORGO and 30.2% on UASpeech, respectively.

Index Terms: Dysarthric speech processing, raw phase and magnitude spectra, single- and multi-stream acoustic modelling

1. Introduction

Dysarthria is the most common neurological speech disorder caused by a disruption in the neuro-motor interface [1]. Due to reduced motor control of the speech articulators, dysarthric speech is often characterised by slower speaking rate, heavily slurred speech, abnormal pauses and repetitions. It reduces the intelligibility of the speech and consequently degrades the performance of the automatic dysarthric speech *recognition* (ADSR) systems, making interaction with the voice-enabled machines much more challenging. Along with ADSR, dysarthric speech *detection* (e.g., [2, 3]) and severity level *classification* (e.g., [4, 5]) are two other active research areas in this context. These three tasks, namely dysarthric speech recognition, detection and classification collectively contribute towards developing more reliable dysarthric speech processing systems which are greatly desirable in healthcare applications and can substantially improve the daily life of people with dysarthria.

The mainstream ASR systems designed for typical speech perform poorly on dysarthric speech. The large systematic mismatch between dysarthric and typical speech, high intra- and inter-speaker variabilities, and data scarcity are three major challenges in ADSR. To enhance the performance of ADSR systems, previous studies have employed data augmentation, (e.g., speed perturbation [6, 7] and voice conversion [8]), high-level speech representations (e.g., bottleneck [9] and autoencoder bottleneck [10] features) and multi-modal representations incorporating features from other modalities (e.g., visual [11] and articulatory [12, 13]).

Raw signal representations such as raw waveform [14–17], raw magnitude [18], raw phase [19], raw real and imaginary parts [20] and, raw source and filter components [21] have been recently applied in acoustic modelling for typical speech. Compared with the task-blind hand-crafted features such as MFCC, the raw representations are richer information-wise. Note that a well-trained deep neural network (DNN) can only effectively process the information and cannot compensate for non-redundant information lost along the engineered front-ends. That is, if task-useful non-redundant information is lost during a hand-crafted feature extraction pipeline, it cannot be recovered by the back-end. When raw signal representations are applied, the information filtering is entirely learned. This minimises the possibility of suboptimal task-useful information loss in the front-end. Such extra information offered by raw signal representations (e.g., raw waveform [22] or raw source-filter [23, 24]) has been shown to be useful in the context of ADSR, too.

Building on [25] where raw phase-based acoustic models for typical speech have been successfully constructed, we explore the effectiveness of the raw phase spectrum in ADSR. The phase and its derivative, namely group delay (GD), have been employed in feature extraction for typical speech recognition, including Modified GD (MGD) [26], Chirp GD [27], Product Spectrum (PS) [28] and parametric GD [29, 30]. The phase spectrum was also utilised for dysarthric speech detection [31]. Nevertheless, its usefulness in the context of ADSR is under-explored, amounting to one study [32] where MGD and PS based cepstral coefficients were compared with MFCC.

In this paper we investigate the usefulness of the raw phase and magnitude spectra for dysarthric speech recognition, detection and classification. Phase spectrum has a sophisticated structure which complicates designing an effective hand-crafted pipeline to capture and represent its information. Raw phase based acoustic modelling leverages the capabilities of DNNs and can effectively extract the information encoded in the phase via learning a fully-trainable pipeline. Further, this framework minimises the possibility of suboptimal task-useful information loss which inadvertently occurs along the engineered pipelines.

To this end, we explore the efficacy of different single- and multi-stream architectures to combine the phase and magnitude-based representations and investigate their training dynamics in terms of cross entropy (CE) loss and WER vs epoch. Experiments are carried out on the TORGO and UASpeech and results are reported for various dysarthric severity levels.

Having described the architecture of the acoustic models in Section 2, we explain the experimental setup in Section 3. Section 4 presents the experimental results along with discussion and Section 5 concludes the paper.

[†] Equal contribution.

Supported by EPSRC Project EP/R012180/1 (SpeechWave).

10.21437/Interspeech.2023-222

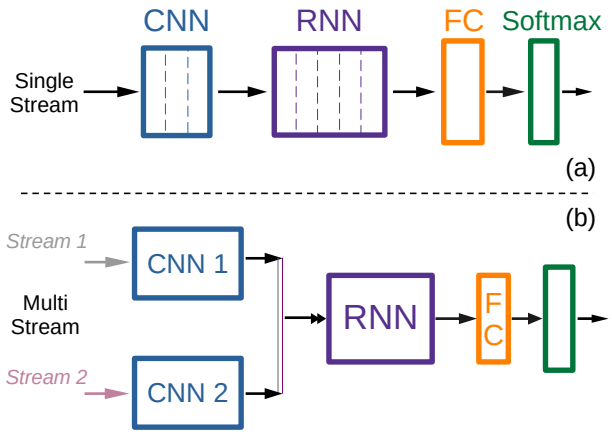

Figure 1: *Employed architectures for acoustic modelling consisting of convolutional (CNN), recurrent and fully-connected (FC) layers. (a) Single-stream, (b) Multi-stream.*

2. Proposed Systems

Fig. 1 illustrates the single and multi-stream architectures, deployed for raw phase and magnitude based acoustic modelling. The single-stream model consists of a cascade of convolutional, recurrent and fully-connected (FC) sub-networks. The convolutional layers extract task-correlated feature maps to effectively represent the input, the recurrent block models the temporal/sequential context (dynamics) and the fully-connected part extracts further abstraction and enhances linear separability to facilitate linear classification via the softmax output layer.

The multi-stream architecture fuses the input streams after pre-processing each stream via individual convolutional blocks. We fuse the pre-processed streams at a medium level, right after the CNN blocks [18]. Note that fusion at higher levels increases the number of trainable parameters (#Params) and fusion at the input level prevents individual pre-processing for each stream. Fusion in the medium level provides an optimal trade-off between pre-processing per stream and keeping #Params within an effectively trainable range.

We have studied two types of feature combination for multi-stream processing: raw phase along with raw magnitude (Mag) or filterbank (FBank) and, duplicating the input features. By duplicating the input e.g., Mag+Mag, we wish to explore the usefulness of increasing the redundancy as well as possibility of benefiting from different initialisations. That is, Mag includes both source and filter components; by passing Mag through parallel CNN blocks initialised differently, hypothetically one stream might focus on the filter (vocal tract) component and the other one on some complementary information such as the source (excitation) part.

3. Experimental Setup

Experiments were conducted on the TORGO [35] and UASpeech [36] dysarthric speech datasets. TORGO includes 15 speakers (eight speakers with three degrees of dysarthria and seven typical speakers) and contains 21 hours (7.3 hours dysarthric and 13.7 hours typical) speech data. UASpeech comprises of 102.7 hours of speech recorded from 29 speakers (16 speakers with dysarthria and 13 typical speakers). The networks were trained using PyTorch-Kaldi [37, 38]. Alignments were taken from the respective Kaldi standard recipe [39]. The CNN sub-network consists of 1-D convolutions with 1 or 3 layers, referred to as 1-L and 3-L, respectively. Configuration of convolutional layers of the raw spectral features in terms of number-of-kernels/kernel-size/pooling-size for raw spectral inputs is as follows: 128/129/3 for 1-L and 128/129/3, 60/5/3 and 60/5/3 for 3-L. The recurrent block consists of a stack of five bidirectional LiGRU [40] layers with 550 nodes in each direction. The fully-connected layer has 1024 ReLU units. Dimensions of the FBank and raw spectral features are 83 (80 FBank + 3 pitch) and 257, respectively.

For data augmentation, we applied speed perturbation using 0.9 (slower), 1.0 (original) and 1.1 (faster) speed change factors, increasing the data by three folds. A 5-fold cross-training TORGO setup was applied for training. The total dataset was divided into five folds with allocating 70% to training, 10% to dev and 20% to test [10]. Experimental results in Tables 1 and 2 are reported in terms of mean±standard deviations over five folds. An independent 200k vocabulary size LibriSpeech trigram language model was employed for decoding [7].

4. Experimental Results and Discussion

Tables 1 and 2 present the WER of different single and multi-stream systems on TORGO, with 1-L and 3-L CNNs. The average WERs are shown for typical and dysarthric speech with different severity levels: mild, moderate (Mod) and severe (Sev).

4.1. Single-stream ADSR systems

Comparing the performance of various features in the single-stream setup (Table 1) shows that the minimum-phase (MinPhase) phase [19] based system achieves comparable results to the raw magnitude based system and outperforms the FBank system on both dysarthric and typical speech. In fact, it returns the best performance among all single-stream systems on typical speech (WER: 9.1%), and results in a highly competitive 31.2% WER on dysarthric speech. The highest performance on dysarthric speech belongs to the raw magnitude based system (WER: 30.4%). It achieves the second best performance on typical speech (WER: 9.7%).

Although the WER of the raw wrapped phase based system is high, it is still remarkably better than a random guess, particularly when 1-L CNN is used. It reflects the relative success of the model in deciphering the wrapped phase despite the data scarcity and complexity of the TORGO task. Note that unwrapping the phase leads to a poorer performance than the wrapped phase. A similar observation was made and discussed in [19].

Across all features, 1-L CNN significantly outperforms the 3-L CNN. For example, in case of the *Mag* system, replacing the 3-L with 1-L CNN, results in 5.3% and 1.4% absolute WER reduction for dysarthric and typical speech, respectively. For the raw *MinPhase* system, the gain is even larger: 7.5% for dysarthric and 3.1% for typical speech.

Note that #Params of the 3-L CNN system is fewer than its 1-L counterpart. That is, although #Params of the CNN block in the 3-L architecture is larger, size of the (hypothetical) flattening layer right after the CNN block and before the first recurrent layer is notably smaller owing to using fewer filters in the third convolutional layer (60 filters). This leads to an architecture with significantly less parameters.

4.2. Multi-stream ADSR systems

As mentioned in Section 2, to build multi-stream systems, two approaches were explored: first, simply duplicating the same

Table 1: *WER of single-stream ADSR systems averaged over various severity levels (Mild, Mod: moderate, Sev: severe) on TORGO.*

Model	**3-L CNN**						**1-L CNN**					
Feature (Single-stream)	**#Params** (Millions)	Severity degrees **Sev**	**Mod**	**Mild**	Average **Dys**	**Typ**	**#Params** (Millions)	Severity degrees **Sev**	**Mod**	**Mild**	Average **Dys**	**Typ**
FBank	10.1	57.4	44.0	15.8	43.3±5.1	14.3±1.9	11.6	48.4	30.8	10.3	34.4±2.0	10.7±0.7
Mag	9.8	48.1	32.4	11.2	35.7±3.4	11.1±1.2	15.6	42.6	27.1	9.6	**30.4**±2.8	9.7±0.3
Wrapped-Phase	9.8	106.0	98.4	98.4	102.5±6.5	98.3±3.9	15.6	75.3	64.9	32.0	61.4±5.6	37.8±5.2
Unwrapped-Phase	9.8	88.4	86.7	64.3	81.5±3.9	68.7±3.6	15.6	81.0	73.5	42.1	68.8±1.9	46.0±4.2
MinPhase	9.8	53.1	37.1	12.4	38.7±3.6	12.2±1.6	15.6	43.8	28.1	9.1	31.2±3.0	**9.1**±0.5

Table 2: *WER of multi-stream ADSR systems averaged over various severity levels (Mild, Mod: moderate, Sev: severe) on TORGO.*

Model	**3-L CNN**						**1-L CNN**					
Feature (Multi-stream)	**#Params** (Millions)	Severity degrees **Sev**	**Mod**	**Mild**	Average **Dys**	**Typ**	**#Params** (Millions)	Severity degrees **Sev**	**Mod**	**Mild**	Average **Dys**	**Typ**
Mag+Mag	10.1	47.0	31.1	9.5	33.5±3.1	9.7±0.7	21.6	44.1	27.9	9.3	31.3±1.9	9.1±0.2
Mag+Mag+Mag	10.3	46.9	30.8	9.4	33.4±1.6	9.6±0.7	27.7	44.0	28.7	9.0	**31.1±2.9**	9.2±0.5
MinPhase+MinPhase	10.1	49.2	32.8	10.9	35.4±3.5	10.3±1.6	21.6	45.3	29.2	9.4	32.2±2.8	9.4±0.5
MinPhase+MinPhase+MinPhase	10.3	48.2	30.0	10.5	34.2±2.6	9.9±0.9	27.7	45.6	30.7	9.7	32.7±3.1	9.9±0.3
FBank+Cos(Phase)	10.2	51.2	33.8	12.1	36.9±3.2	11.5±1.1	17.7	50.8	35.9	11.5	37.0±1.4	12.1±0.9
FBank+MinPhase	10.2	47.3	30.9	11.5	34.2±2.6	10.6±1.0	17.7	44.7	29.1	10.1	32.1±2.6	10.4±0.6
FBank+Mag	10.2	46.8	30.5	10.8	33.6±2.1	10.6±0.9	17.7	43.7	28.1	10.2	31.4±2.4	10.0±0.7
Mag+WrappedPhase	10.1	47.9	31.5	10.1	34.2±3.0	10.0±1.9	21.6	48.4	33.7	10.9	35.2±2.1	10.9±1.5
Mag+Cos(Phase)	10.1	47.3	29.6	9.8	33.7±1.6	9.8±0.3	21.6	48.3	34.1	10.2	35.0±1.7	10.8±0.5
Mag+Sin(Phase)	10.1	48.9	30.8	10.2	34.7±3.4	10.1±0.8	21.6	48.6	32.4	10.2	34.8±2.5	10.5±0.4
Mag+MinPhase	10.1	48.4	31.7	10.5	34.6±2.6	10.4±1.5	21.6	44.2	28.4	9.2	31.4±2.5	**9.0**±0.2

Table 3: *Comparison with other ADSR systems on TORGO.*

	Mag	**MinPhase**	**Mag+MinPhase**	**[33]**	**[24]**	**[20]**
Dys	30.4	31.2	31.4	40.7	33.1	31.7
Typ	9.7	9.1	9.0	-	10.3	10.2

Table 4: *WER of various ADSR systems on UASpeech.*

Feature	**FBank**	**Mag**	**MinPhase**	**Mag+MinPhase**	**[34]**
UASpeech	31.7	30.4	30.8	**30.2**	30.5

features to feed multiple input streams, motivated by the possibility of taking advantage of extra redundancy and capturing complementary information. For example, *Mag+Mag* and *Mag+Mag+Mag* in Table 2 refers to multi-stream systems with two and three input streams, respectively, where all inputs are raw magnitude spectrum. Second, using different features as input streams, e.g., *Mag+Cos(phase)* means the input streams are raw magnitude and cosine of the wrapped phase.

Comparing Table 2 and Table 1 shows that multi-streaming by replicating the input feature leads to consistent performance gain only when the 3-L CNN architecture is used. In this case, the WER of the *Mag+Mag* system is 33.5% and 9.7% for dysarthric and typical speech, lower than its single-stream *Mag* counterpart by 2.2% and 1.4% (absolute), respectively. However, when 1-L CNN is employed, the performance gets better only for typical speech by 0.5% (absolute). Also note that there is a diminishing return after replicating the input stream three times (*Mag+Mag+Mag*). Similar observations can be made for the phase-based *MinPhase*, *MinPhase+MinPhase* and *MinPhase+MinPhase+MinPhase* systems.

Combining FBank with the raw phase-based representations indicates that while coupling cos(phase) with FBank leads to a poorer performance, *FBank+MinPhase* results in 2.3% and 0.4% absolute WER reductions (relative to FBank, with 1-L CNN) on dysarthric and typical speech, respectively.

Reciprocally, if we consider the *MinPhase* system as the baseline, adding FBank degrades the performance. This rather surprising observation, however, does not mean FBank feature is not informative but implies it does not offer extra information useful to the task. FBank is a lossy representation of the raw magnitude spectrum with finer sampling at low frequencies and filters mimicking the spectral masking (e.g., triangular filters in MFCC or trapezoidal in PLP). Given DNNs' capabilities in learning representative patterns, these knowledge-based transformations do not appear to be critically needed.

The effect of adding various raw phase-based representations to the raw magnitude spectrum is shown in the last part of Table 2. Such combinations lead to a consistent performance gain for both dysarthric and typical speech when 3-L CNN is used. The largest improvement belongs to fusing the cosine of the unwrapped phase, cos(phase), with the raw magnitude.

After applying 1-L CNN, fusing the raw magnitude and phase-based representations does not improve the performance. We hypothesise this is owing to a remarkable increase in the model parameters when moving from single to multi-stream architecture, in which case, the amount of training data might not be sufficient for realising the full potential of such systems.

Finally, to put the reported numbers in context, Table 3 demonstrates the performance of the proposed systems along with previous studies on TORGO [20, 24, 33].

In the next stage, we evaluated our best-performing multi-stream ADSR systems on another widely used dysarthric speech corpus, namely UASpeech. As seen in Table 4, the *Mag+MinPhase* system achieves the best WER (30.2%), outperforming both *Mag* and the *MinPhase* systems as well as the previous highly competitive work [34].

Table 5: *Accuracy (%) of dysarthric speech detection and severity classification. WAvg: weighted average.*

Task	Classification (severity degrees)					Detection		
Feature	**Sev**	**Mod**	**Mild**	**Typ**	**WAvg**	**Dys**	**Typ**	**WAvg**
FBank	90.5	78.7	78.3	76.7	81.3	85.1	88.3	86.7
Mag	92.3	86.7	85.4	77.0	**86.0**	87.1	89.2	**88.1**
MinPhase	91.4	81.3	81.2	73.3	81.8	85.3	86.7	85.8
Mag+MinPhase	91.6	85.8	82.5	73.3	83.8	87.0	88.5	87.7

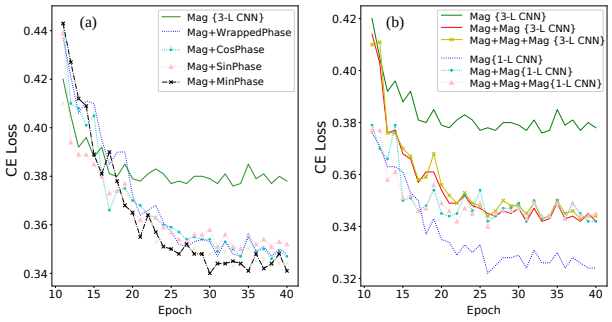

Figure 2: *CE vs. epoch for the single and multi-stream acoustic models on TORGO. (a) 3-L CNNs, (b) 1-L and 3-L CNNs.*

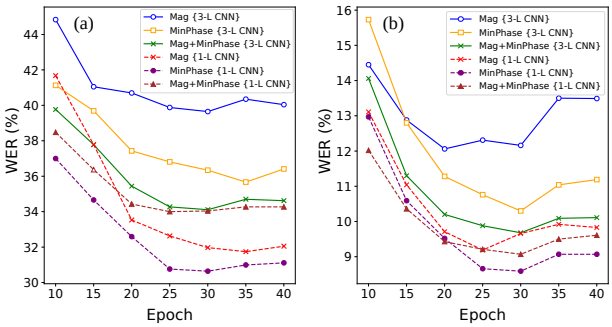

Figure 3: *WER vs. epoch for the single and multi-stream acoustic models on TORGO. (a) Dysarthric, (b) Typical.*

4.3. Training dynamics

We also explore the training dynamics of models in terms of performance metrics vs. epoch. Fig. 2 shows the CE loss vs. epoch on the dev set for various systems. As seen, adding phase-based representations improves the loss on systems with a 3-L CNN. Comparing the *Mag* with *Mag+Mag* systems shows that when the 3-L CNN is used, multi-streaming reduces CE and slows down the convergence. In contrast, on architectures with a 1-L CNN, it leads to larger CE and faster convergence. Also note that, the 1-L CNN systems require more epochs for convergence than their 3-L counterparts. For example, the knee points of the 1-L and 3-L single-stream *Mag* systems are around 25 and 20 epochs, respectively. This can be explained considering the fact that the 1-L CNN system has a larger #Params (as discussed in 4.1 and shown in Tables 1 and 2).

Fig. 3 illustrates the training dynamics in terms of WER vs. epoch for the phase and magnitude-based systems. Compared with the FBank and raw magnitude based systems, the MinPhase based single-stream system requires more training steps to converge. The more intricate structure of the phase spectrum makes the learning process slower and can explain this observation. Lastly, for typical speech, training with more than 30 epochs barely improves the performance, while dysarthric speech can still benefit from further training. This can be attributed to the fact that dysarthric speech is more complex to model which slows down learning the underlying patterns.

4.4. Detection and classification

We also investigated the effectiveness of utilising the raw magnitude and phase spectra in the dysarthric speech detection and classification tasks. The former involves a binary classification between dysarthric and typical speech, while the latter aims to categorise the speech into one of four severity levels: severe (Sev), moderate (Mod), mild, or typical.

To this end, we dumped the CNN feature maps and employed the Support vector machines (SVM) [41] with the Radial basis function (RBF) kernel for both two-class detection and four-class classification tasks. SVMs were trained using scikit-learn toolkit [42] and for four-class classification we used one-vs-rest strategy.

Table 5 presents the performance of the proposed systems in terms of weighted accuracy. We first calculated the accuracy per class (recall) and then computed the weighted average where the weights are the class prior probabilities in each task.

As can be seen, the raw magnitude based system consistently outperforms others in both detection and classification tasks, achieving 86.0% and 88.1% average accuracy. This feature representation works especially well in classifying moderate to severe dysarthric speech. It is also notable that the minimum-phase phase spectra achieve comparable classification performance to FBank on dysarthric speech.

5. Conclusion

In this paper, we investigated and demonstrated the efficacy of automatic dysarthric speech recognition, detection and classification using raw phase and magnitude spectra. Our acoustic models were a cascade of convolutional, recurrent and fully-connected layers, with single- and multi-stream architectures. Experiments were carried out on the widely-used TORGO and UASpeech dysarthric speech databases, and results have been reported for various severity degrees. We explored various configurations along with analysing the training dynamics (CE loss and WER vs. epoch) of different models. Our best systems achieved a highly competitive performance of 30.4% and 9.0% WERs on dysarthric and typical speech on TORGO and 30.2% WER on UASpeech. Future work includes using raw phase and magnitude spectra in constructing multi-task systems jointly trained to perform recognition, detection and classification tasks within a unifying model. This will leverage the synergies between these tasks and is a broad avenue for future work.

6. References

[1] J. Duffy, *Motor speech disorders e-book: Substrates, differential diagnosis, and management*. Elsevier Health Sciences, 2019.

[2] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's disease," in *INTERSPEECH*, 2015, pp. 95–99.

[3] J. Tracy, Y. Özkanca, D. Atkins, and R. Ghomi, "Investigating voice as a biomarker: deep phenotyping methods for early detection of parkinson's disease," *Journal of Biomedical Informatics*, vol. 104, p. 103362, 2020.

[4] A. Joshy and R. Rajan, "Automated dysarthria severity classification: A study on acoustic features and deep learning techniques," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1147–1157, 2022.

[5] A. A. Joshy and R. Rajan, "Dysarthria severity classification using multi-head attention and multi-task learning," *Speech Communication*, vol. 147, pp. 1–11, 2023.

[6] S. Liu, S. Hu, X. Xie, and H. Meng, "Recent progress in the cuhk dysarthric speech recognition system," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[7] Z. Yue, F. Xiong, H. Christensen, and J. Barker, "Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition," in *ICASSP*, 2020.

[8] W. Huang, B. Halpern, L. Violeta, O. Scharenborg, and T. Toda, "Towards identity preserving normal to dysarthric voice conversion," in *ICASSP*, 2022.

[9] V. Yılmaz, E.and Mitra, G. Sivaraman, and H. Franco, "Articulatory and bottleneck features for speaker-independent asr of dysarthric speech," *Computer Speech and Language*, vol. 58, pp. 319–334, 2019.

[10] Z. Yue, H. Christensen, and J. Barker, "Autoencoder bottleneck features with multi-task optimisation for improved continuous dysarthric speech recognition," in *INTERSPEECH*, 2020.

[11] E. Salama, R. El-Khoribi, and M. Shoman, "Audio-visual speech recognition for people with speech disorders," *International Journal of Computer Applications*, vol. 96, no. 2, 2014.

[12] F. Xiong and H. Barker, J.and Christensen, "Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition," in *ITG-Symposium*, 2018.

[13] Z. Yue, E. Loweimi, Z. Cvetkovic, H. Christensen, and J. Barker, "Multi-modal acoustic-articulatory feature fusion for dysarthric speech recognition," in *ICASSP*, 2022.

[14] M. Ager, Z. Cvetković, and P. Sollich, "Combined waveform-cepstral representation for robust speech recognition," in *ISIT*, 2011.

[15] J. Yousafzai, P. Sollich, Z. Cvetkovic, and B. Yu, "Combined features and kernel design for noise robust phoneme classification using support vector machines," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 5, pp. 1396–1407, 2010.

[16] M. Ravanelli and Y. Bengio, "Speaker and speech recognition from raw waveform with SincNet," in *ICASSP*, 2019.

[17] E. Loweimi, P. Bell, and S. Renals, "On learning interpretable CNNs with parametric modulated kernel-based filters," in *INTERSPEECH*, 2019.

[18] ——, "Raw sign and magnitude spectra for multi-head acoustic modelling." in *INTERSPEECH*, 2020.

[19] E. Loweimi, Z. Cvetkovic, P. Bell, and S. Renals, "Speech acoustic modelling from raw phase spectrum," in *ICASSP*, 2021.

[20] E. Loweimi, Z. Yue, P. Bell, S. Renals, and Z. Cvetkovic, "Multi-stream acoustic modelling using raw real and imaginary parts of the fourier transform," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 876–890, 2023.

[21] E. Loweimi, Z. Cvetkovic, P. Bell, and S. Renals, "Speech Acoustic Modelling Using Raw Source and Filter Components," in *INTERSPEECH*, 2021.

[22] Z. Yue, E. Loweimi, H. Christensen, J. Barker, and Z. Cvetkovic, "Dysarthric speech recognition from raw waveform with parametric CNNs," in *INTERSPEECH*, 2022.

[23] Z. Yue, E. Loweimi, and Z. Cvetkovic, "Raw source and filter modelling for dysarthric speech recognition," in *ICASSP*, 2022.

[24] Z. Yue, E. Loweimi, H. Christensen, J. Barker, and Z. Cvetkovic, "Acoustic modelling from raw source and filter components for dysarthric speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2968–2980, 2022.

[25] E. Loweimi, Z. Cvetkovic, P. Bell, and S. Renals, "Speech acoustic modelling from raw phase spectrum," in *ICASSP*, 2021.

[26] H. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *ICASSP*, 2003.

[27] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," *Speech Communication*, vol. 49, no. 3, pp. 159 – 176, 2007.

[28] D. Zhu and K. Paliwal, "Product of power spectrum and group delay function for speech recognition," in *ICASSP*, vol. 1, May 2004, pp. I–125–8 vol.1.

[29] E. Loweimi and S. M. Ahadi, "A new group delay-based feature for robust speech recognition," in *ICEM*, 2011, pp. 1–5.

[30] E. Loweimi, S. Ahadi, and T. Drugman, "A new phase-based feature representation for robust speech recognition," in *ICASSP*, 2013.

[31] P. Janbakhshi and I. Kodrasi, "Experimental investigation on stft phase representations for deep learning-based dysarthric speech detection," in *ICASSP*, 2022.

[32] S. Sehgal, S. Cunningham, and P. Green, "Phase-based feature representations for improving recognition of dysarthric speech," in *SLT*, 2018.

[33] T. Mariya Celin, P. Vijayalakshmi, and T. Nagarajan, "Data augmentation techniques for transfer learning-based continuous dysarthric speech recognition," *Circuits, Systems, and Signal Processing*, vol. 42, no. 1, pp. 601–622, 2023.

[34] D. Wang, J. Yu, X. Wu, L. Sun, X. Liu, and H. Meng, "Improved end-to-end dysarthric speech recognition via meta-learning based model re-initialization," in *ISCSLP*. IEEE, 2021, pp. 1–5.

[35] F. Rudzicz, A. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[36] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[37] M. Ravanelli, T. Parcollet, and Y. Bengio, "The pytorch-kaldi speech recognition toolkit," in *ICASSP*, 2019.

[38] A. Paszke *et al.*, "Automatic differentiation in PyTorch," in *NIPS Workshop on Autodiff*, 2017, pp. 1–4.

[39] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[40] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 92–102, 2018.

[41] N. Cristianini, J. Shawe-Taylor *et al.*, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[42] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.