INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Phase-Aware Signal Processing for Automatic Speech Recognition

Johannes Fahringer, Tobias Schrank, Johannes Stahl, Pejman Mowlaee, Franz Pernkopf

Signal Processing and Speech Communication Lab
Graz University of Technology, Graz, Austria
{johannes.fahringer,tobias.schrank,johannes.stahl,pejman.mowlaee,pernkopf}@tugraz.at

Abstract

Conventional automatic speech recognition (ASR) often neglects the spectral phase information in its front-end and feature extraction stages. The aim of this paper is to show the impact that enhancement of the noisy spectral phase has on ASR accuracy when dealing with speech signals corrupted with additive noise. Apart from proof-of-concept experiments using clean spectral phase, we also present a phase enhancement method as a phase-aware front-end and modified group delay as a phase-aware feature extractor, and the combination thereof. In experiments, we demonstrate the improved performance for each individual component and their combination, compared to the conventional phase-unaware Mel Frequency Cepstral Coefficients (MFCCs)-based ASR. We observe that the estimated phase information used in the front-end or feature extraction component improves the ASR word accuracy rate (WAR) by 20.98 % absolute for noise corrupted speech (averaged over SNRs ranging from 0 to 20 dB).

Index Terms: Phase estimation, automatic speech recognition, modified group delay features, Mel Frequency Cepstral Coefficient (MFCC).

1. Introduction

The phase spectrum is known to be a controversial topic. Several early studies point out the unimportance [1] and importance [2–4] of spectral phase information from a perceptual viewpoint. Several more recent studies reported positive impact of spectral phase in different speech processing applications including speech enhancement [5–8], speech intelligibility prediction [9, 10] and ASR [11] (see [12, 13] for an overview). Conventional automatic speech recognition (ASR) systems often rely solely on the magnitude spectrum and are built upon short-time amplitude-derived features [14]. The common practice is to discard the spectral phase information and feed ASR with features derived from the power spectrum.

However, in the last two decades plenty proposals have been made to apply some sort of phase-aware signal processing for ASR. These studies can be divided into two categories: front-end (Fr) signal processing and feature extraction (Fx), as shown in Figure 1. In front-end processing, phase-aware (PA) enhancement-schemes are applied to noisy speech. As some examples, we refer to complex spectral subtraction [15] and phase-sensitive filters (PSF) learned by deep neural network (DNN) [16]. Furthermore, there have been several studies that used group delay and modified group delay (MODGD) as features for ASR, obtaining good results [11, 17]. Group delay (GD) features show robust performance in noise [18] and are capable to represent the speech formants at a high resolution [19]. Inspired by the convincing results regarding phase-aware signal processing in the front-end [15, 16] or feature extraction [17, 20, 21] of ASR, we present a systematic study, addressing two questions: 1) can phase enhancement (PE) contribute to improved ASR results when combined with the noisy spectral magnitude for signal reconstruction (Figure 1: Fr (PA)) 2) encouraged by improved ASR reported for phase aware features (e.g. MODGD), is it possible to benefit from a phase-enhancement used in combination with phase aware features (Figure 1: Fr (PA)+Fx (PA)).

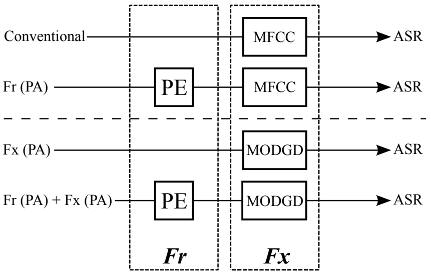

Figure 1: Different architectures for phase-aware processing in ASR, front-end (*Fr*), feature extraction (*Fx*), and both combined compared to conventional phase-unaware MFCC-based ASR (top). We define *PE* as phase enhancement and *PA* as an acronym for phase-awareness.

2. Related Work

2.1. Feature Extraction

The possibility of applying short-time phase spectrum for ASR was studied by Alsteris et al. [11, 22]. Group delay features have been successfully used in the literature for improved ASR performance. As some examples for GD features we refer to model-based [21] and modified GD (MODGD) [17]. In all these studies it has been reported that phase-aware features contribute to improved ASR performance. More recently Loweimi et al. [23] showed that the conventional source-filter model leads to some loss of information regarding the vocal tract, which is known to be useful for ASR. In order to capture the temporal evolution of the vocal tract and the excitation (both are helpful for ASR), a blind deconvolution of speech, based on phase only processing, was proposed. An improved ASR performance up to 8.5% word accuracy rate (WAR) compared to MFCC features was reported.

This work was supported by the Austrian Science Fund (FWF) under the project numbers P28070-N33 and P25803-N15 and the LEAD Project funded by Graz University of Technology.

The K-Project ASD is funded in the context of COMET Competence Centers for Excellent Technologies by BMVIT, BMWFW, Styrian Business Promotion Agency (SFG), the Province of Styria - Government of Styria and Vienna Business Agency. The programme COMET is conducted by Austrian Research Promotion Agency (FFG)

http://dx.doi.org/10.21437/Interspeech.2016-823

2.2. Front-end Processing

Schlüter and Ney combined MFCCs and features purely based on phase information and reported a relative improvement of up to 25 % word error rate (WER) [24]. Golik et al. [25] trained convolutional neural networks (CNNs) with raw time signals to learn acoustic models which is in contrast to conventional feature extraction relying on a filterbank and some energy feature extraction. The CNN is capable of learning the critical band energy filters distributed non-linearly similar to a Mel filterbank. Using raw time signals together with CNNs, a WER comparable to MFCC-based ASR was achieved. The position of the learned filters in time is distributed uniformly within the context [20]. This is in contrast to the MFCC-based features, where the position of the filters in time is bounded to the center of the stacked samples of the signal. As these time offsets or shifts are only represented by the phase spectrum, the authors concluded that a neural network is capable of learning different filters at different segments of the signal. By providing raw time signals to a neural network, the acoustic models can learn non-stationary patterns with a reduced loss of information in comparison to processing used in common feature extraction. Kleinschmidt et al. [15] reinforced complex spectral subtraction and showed improved ASR for scenarios in which the clean phase is available. Finally, a DNN was presented with a PSF where the cosine of phase difference between the spectral clean phase and noisy phase was taken into account [16]. The PSF was used as the mapping function to be learned by the DNN. Improved automatic speech recognition performance was reported for PSF compared to phase-insensitive mask functions (e.g., Wiener filter).

3. Phase-only Speech Enhancement

3.1. Proposed Phase Estimator

In this work, we propose a phase estimator as a front-end processor of the noise corrupted speech. We formulate the estimator in the *short time Fourier transform* (STFT) domain, starting from the following signal model under the assumption of additive noise

$$\underbrace{R(k,l)\exp\left(j\vartheta(k,l)\right)}_{Y(k,l)} = \underbrace{A(k,l)\exp\left(j\alpha(k,l)\right)}_{X(k,l)} + D(k,l), \tag{1}$$

where $Y(k,l)$ denotes the noisy speech coefficients with absolute value $R(k,l)$ and phase $\vartheta(k,l)$ and $D(k,l)$ is the noise STFT coefficient at frequency bin k and frame index l, respectively. The clean speech coefficients $X(k,l)$ consist of absolute value $A(k,l)$ and phase $\alpha(k,l)$. We obtain the corresponding time domain signals by applying the *inverse short time Fourier transform* (iSTFT$(\cdot)$). For the sake of notational simplicity we drop the frequency index k and the frame index l wherever possible. Similar to [26] we formulate a *maximum a posteriori* (MAP) criterion under the assumption that the true spectral phase follows a von Mises distribution around a prior phase estimate denoted by α_μ. However, in contrast to [26], here, we consider the noisy STFT phase directly rather than the phase of one sinusoid in noise. The von Mises distribution of a phase α is characterized by its mean value α_μ and its concentration parameter κ. The mean value α_μ could be obtained from any phase-estimator, e.g. [27] (for an overview see [6]). The corresponding probability density function (pdf) is given by

$$p(\alpha) = \frac{\exp\left(\kappa\cos\left(\alpha-\alpha_\mu\right)\right)}{2\pi I_0\left(\kappa\right)}, \tag{2}$$

where $I_\nu\left(\cdot\right)$ denotes the modified Bessel function of the first kind and order ν. Further, if we assume the real and imaginary parts of the noise coefficients to be independently Gaussian distributed with zero mean and variance $\frac{\sigma_d^2}{2}$, the pdf of the noisy observation Y conditioned on the clean speech coefficient is given by

$$p\left(Y|A,\alpha\right) = \frac{1}{\pi\sigma_d^2}\exp\left(-\frac{|Y-A\exp\left(j\alpha\right)|}{\sigma_d^2}\right). \tag{3}$$

In the following, we seek to maximize the posterior probability $p\left(A,\alpha|Y\right)$. However, since we do not have access to the true values of A and α by employing Bayes' theorem and assuming independency of amplitude and phase we obtain

$$\begin{aligned}\hat{\alpha}_{\text{MAP}} &= \arg\max_{\alpha}\frac{p\left(Y|A,\alpha\right)p\left(A\right)p\left(\alpha\right)}{p\left(Y\right)}\\ &= \arg\max_{\alpha} p\left(Y|A,\alpha\right)p\left(\alpha\right).\end{aligned} \tag{4}$$

By taking the derivative of the logarithm of the posterior and setting it to zero we obtain the MAP phase estimate $\hat{\alpha}_{\text{MAP}}$.

$$\begin{aligned}0 &\overset{!}{=} \frac{\partial\log\left(p\left(Y|A,\alpha\right)p\left(\alpha\right)\right)}{\partial\alpha}\Big|_{\alpha=\hat{\alpha}_{\text{MAP}}},\\ 0 &= \frac{2AR}{\sigma_s^2}\sin\left(\hat{\alpha}_{\text{MAP}}-\vartheta\right)-\kappa\sin\left(\hat{\alpha}_{\text{MAP}}-\alpha_\mu\right),\\ \hat{\alpha}_{\text{MAP}} &= \arctan\left(\frac{\frac{2RA}{\sigma_d^2}\sin\left(\vartheta\right)+\kappa\sin\left(\alpha_\mu\right)}{\frac{2RA}{\sigma_d^2}\cos\left(\vartheta\right)+\kappa\cos\left(\alpha_\mu\right)}\right).\end{aligned} \tag{5}$$

Since the parameters of the von Mises distribution, κ and α_μ, are not known a priori, they need to be estimated from the noisy observation, explained in the following. We assume the noise phase to be zero-mean. Under this assumption we obtain the parameters κ and α_μ of the von Mises distribution by calculating the sample mean of consecutive coefficients which is given as follows [28]

$$z(k,l_{\mathcal{VM}}) = \frac{1}{|\mathcal{L}|}\sum_{l'\in\mathcal{L}} R(k,l')\exp\left(j\vartheta\left(k,l'\right)\right), \tag{6}$$

where $l_{\mathcal{VM}}$ is the frame index resulting from the frameshift $S_{\mathcal{VM}}$ used to obtain the von Mises parameters. $\mathcal{L}$ is the set of frames around the center frame $l_{\mathcal{VM}}$ for which we assume the speech signal to be stationary and $|\mathcal{L}|$ denotes the number of frames l' within the set $\mathcal{L}$. The mean phase estimate is given by

$$\hat{\alpha}_{\mu,\mathcal{VM}}(k,l_{\mathcal{VM}}) = \angle z(k,l_{\mathcal{VM}}), \tag{7}$$

and the maximum likelihood estimate of the concentration parameter $\kappa(k,l_{\mathcal{VM}})$ is obtained via the mean resultant length $|z(k,l_{\mathcal{VM}})|$, i.e.,

$$\hat{\kappa}_{\mathcal{VM}}(k,l_{\mathcal{VM}}) = f^{-1}(|z(k,l_{\mathcal{VM}})|), \quad \text{with} \quad f(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}. \tag{8}$$

There exist several approximations of $f^{-1}(\cdot)$ [28].

3.2. Signal Reconstruction

Plugging (8) and (7) into (5) yields the MAP phase estimate, which is used to reconstruct the time domain *phase enhanced* signal given by

$$\hat{x}_{\mathrm{PE}}(n) = \mathrm{iSTFT}\left(R \exp\left(j\hat{\alpha}^{\mathrm{MAP}}\right)\right). \tag{9}$$

In order to evaluate the upper performance bounds due to the front-end processing unit we also consider a *clean phase* scenario, determined by the following synthesis equation:

$$\hat{x}_{\mathrm{CP}}(n) = \mathrm{iSTFT}\left(R \exp\left(j\alpha\right)\right). \tag{10}$$

4. Experiments

4.1. Experimental Setup

For our experiments we used the GRID corpus [29] as speech material due to its simple pattern grammar. We downsampled all utterances from 16 kHz to 8 kHz so that the audio material has a higher similarity to telephone audio. The STFT framelength of the phase estimator is 32 ms which corresponds to 256 samples at a sampling rate of $f_s = 8\,\mathrm{kHz}$. In order to avoid phase jumps from one frame to the next, we set $S_{\mathcal{VM}} = 1$ sample, yielding $|\mathcal{L}| = 256$ frames. As analysis window we select *Hamming*, which is a typical choice in speech enhancement. Since in a practical scenario we do not know the clean spectral amplitude A in (5), we estimate it using the STSA-MMSE estimator proposed in [30] together with the minimum statistics noise power spectral density (PSD) estimator proposed in [31]. The corresponding implementation was taken from [32]. We chose the frameshift of the amplitude estimator with $S = 16\,\mathrm{ms} \cdot f_s$. In order to have a consistent frame setup, we reduce the number of frames of the von Mises parameters following $\hat{\kappa}(k,l) = \hat{\kappa}_{\mathcal{VM}}(k, S\cdot(l-1)+1)$ and $\hat{\alpha}_{\mu}(k,l) = \hat{\alpha}_{\mu,\mathcal{VM}}(k, S\cdot(l-1)+1)$.
We split the GRID corpus into training, development and test set in exactly the same way as it was proposed for the CHiME 2 challenge (Track 1) [33], i.e. the training set comprises 500 utterances of each of the 34 speakers (18 male, 16 female). The development set and test set contain 600 utterances of every speaker. For the test set – not for the training and development set – we mixed all utterances with two different noise types, namely, train noise and babble noise (representing both, stationary and non-stationary noise types), both taken from the AURORA 2 database [34]. For each utterance and noise type we produced five noisy versions with SNRs ranging from 0 dB to 20 dB SNR in 5 dB steps. In contrast to the second CHiME challenge we refrain from presenting results for 25 dB and 30 dB SNR due to insignificant differences between any processing methods we applied. It is unproblematic to reuse the same noise signal for all utterances in the test set since we only trained on clean audio data. While speech material and composition of our audio material is identical to the second CHiME challenge, the noisy audio files used for testing are not.

4.2. Automatic Speech Recognition

For the experiments performed in this paper we used the ASR baseline system released for the CHiME 2 [33] challenge which is based on the HTK toolkit [35]. In order to verify that our findings also hold for a more modern ASR system, we also performed experiments with a second ASR system based on the Kaldi toolkit [36]. Both systems extract either 13 MFCCs or 13 MODGD features [17, 37] as well as their delta and acceleration coefficients yielding 39 features in total. The Kaldi-based system employs a hybrid DNN architecture. The input features are subject to a series of transformations which have proved to be effective in noisy environments: cepstral mean and variance normalization, linear discriminant analysis, maximum likelihood linear transform, and feature-space maximum likelihood linear regression. The acoustic model is a 8-layer DNN-HMM that is generatively pre-trained using a deep belief network. Each hidden layer contains 1024 neurons. The DNN uses state-level alignments from a GMM-HMM acoustic model which features speaker adaptive training. The DNN is then trained by 4 iterations of sequence-discriminative training employing a state-level minimum Bayes risk criterion. Both ASR systems are evaluated in accordance to the CHiME 2 challenge: Of the seven words in each utterance, only the recognition of letters and digits is relevant for calculating the word accuracy rate.

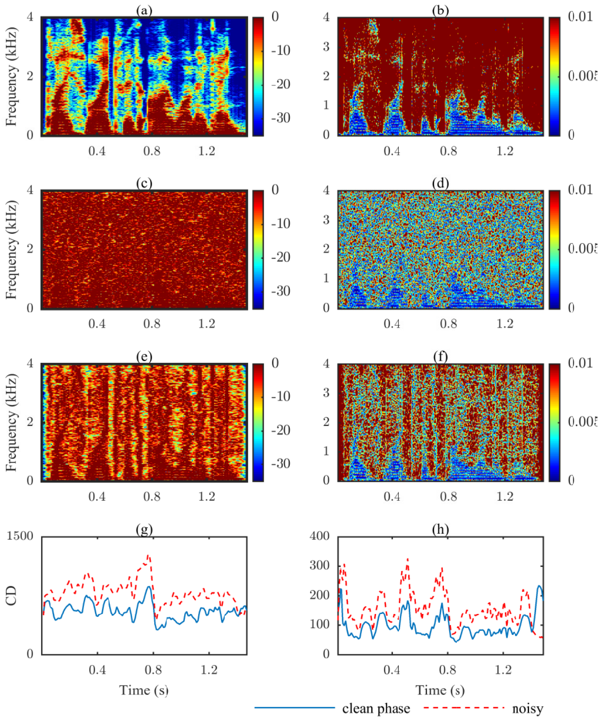

Figure 2: Impact of the clean spectral phase on a noisy signal for the male (Grid Corpus Speaker ID: 27) speech sample *"place white with i one again"*: (a) Spectrogram of the clean signal, (b) MODGD clean signal, (c) Noisy signal (SNR = 0 dB), (d) MODGD of the noisy signal, (e) Noisy signal combined with clean spectral phase (10), (f) MODGD clean phase signal, (g) Cepstral distance (CD) of MFCCs, (h) CD of MODGD

4.3. Cepstral Distance as a Proof of Concept

Our first experiment deals with the question if the enhancement of a noise corrupted signal's spectral phase impacts the extracted features towards their clean pendant. We utilize the clean phase signal obtained from (10) to compare its extracted features with those obtained from the noisy signal (unprocessed). In order to make the enhancement's influence on the features more visible, we calculate the cepstral distance (CD). The CD between the feature vector under test and the

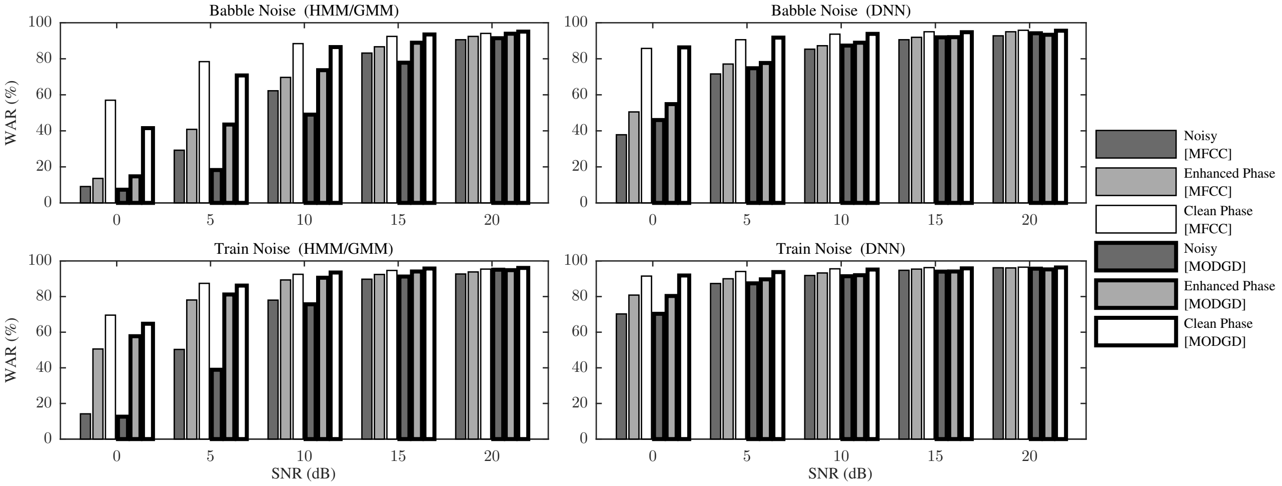

Figure 3: Summarized recognition results. The panels on left hand side show the recognition results for both noise types using the HMM/GMM system and the panels on the right hand side show the results using the DNN system respectively. The results are shown for the noisy signals, the enhanced signals (9) and the clean phase signals (10) for MODGD and MFCC feature extraction.

corresponding clean feature vector is calculated for both feature types. It is defined as the l_2-norm of the difference of the corresponding feature vectors.

Figure 2 illustrates the impact of phase enhancement on a speech signal corrupted with white noise at an SNR of 0 dB. In (e) we picture the magnitude enhancement due to the clean phase (10). The spectral structure of the clean speech signal (a) is partially reconstructed in the enhanced spectrogram (e) while this structure is lost in the spectrogram of the noisy speech (c). Similar observations can be made by examining the modified group delay plots in the right column of Figure 2 (b,d,f). The lower panels show the progression of the CD for MFCC (g) and MODGD (h) features over time. It is apparent that the CD related to clean phase features is in both cases distinctly lower than the CD related to the noisy features. The clean phase signal (10) is more similar to the clean signal than the noisy signal (unprocessed) in terms of CD.

4.4. Speech Recognition Results

For the evaluation of the front-end stage we consider three scenarios:

1. Noisy signal: No enhancement is performed
2. Clean phase signal: Noisy spectral amplitude combined with clean phase, $\hat{x}_{\mathrm{CP}}$, as defined in (10)
3. Phase enhanced signal: Noisy spectral amplitude combined with enhanced phase, $\hat{x}_{\mathrm{PE}}$, as defined in (9)

These scenarios are combined with the described feature extraction stages. The left hand side panels of Figure 3 show the recognition results obtained with the HMM/GMM system. The upper panel shows the WAR for babble noise dependent on the SNR. The enhancement improves the MFCC baseline by 5.76% on average. Although utilizing the MODGD features for the noisy signal leads to a degraded WAR compared to MFCC features, the phase-enhancement boosts the performance of the MODGDs on average by 14.18% which is 2.32% higher than for the MFCC case using the enhancement front-end (PE). The results using the DNN system (Right hand side Panels of Figure 3) show higher overall WAR due to the different feature processing steps and a more powerful acoustic model compared to the HMM/GMM system (see 4.2). The noisy MODGD outperforms the MFCC baseline and the achievable absolute performance gain due to the phase enhancement is similar for MFCCs and MODGD. The ASR performance benefits most from the phase processing for stationary train noise. On average, 20.98% absolute improvement is achieved for the GMM/HMM system when employing both phase enhancement and phase-aware features. The combination of phase enhancement and phase-aware features also leads to significant improvements in WAR for non-stationary babble noise. However, due the speech-like nature of babble noise, phase estimation is considerably more difficult and hence phase-enhancement lead to smaller improvements in comparison to train noise. In comparison to MFCCs, the phase-aware MODGD features improve the WAR for noisy signals by 8.17%, for phase-enhanced signals by 4.33% and for clean phase signals by 0.58% absolute at an SNR of 0 dB babble noise when using the DNN system. We therefore conclude on the two research questions posed in this work.

- Phase-aware front-end processing contributes positively to the performance of conventional MFCC-based ASR.
- An increased performance gain is achievable by combining phase-aware front-end processing with a phase-aware feature extractor.

5. Conclusion

This paper addressed the question how an automatic speech recognition system is affected by phase-aware processing at the front-end and feature extraction stages. For feature extraction we considered MFCCs and modified group delay features as phase-unaware and phase-aware possibilities, respectively. We showed that an estimate of the clean phase combined with the noisy amplitude contributes beneficially to ASR. Throughout various experiments it was revealed that incorporating phase information at front-end or feature extraction stages yields improved ASR performance. A combination of phase-aware front-end and feature extraction showed a large absolute improvement of up to 20.98% WAR compared to the baseline MFCC-based ASR results. The remaining gap between the proposed scheme and the clean phase scenario motivates for further research in the direction of applying phase-aware signal processing for ASR.

6. References

[1] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 30, no. 4, pp. 679–681, 1982.

[2] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, no. 5, pp. 529 – 541, May 1981.

[3] P. Vary, "Noise suppression by spectral magnitude estimation mechanism and theoretical limits," *Signal Processing*, vol. 8, no. 4, pp. 387 – 400, 1985.

[4] L. Li, H. Jialong, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Elsevier speech communication*, vol. 22, no. 4, pp. 403–417, Sept. 1997.

[5] K. K. Paliwal, K. K. Wojcicki, and B. J. Shannon, "The importance of phase in speech enhancement," *Elsevier speech communication*, vol. 53, no. 4, pp. 465–494, 2011.

[6] P. Mowlaee and J. Kulmer, "Phase estimation in single-channel speech enhancement: Limits-potential," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 23, no. 8, pp. 1283–1294, Aug 2015.

[7] J. Kulmer and P. Mowlaee, "Harmonic phase estimation in single-channel speech enhancement using von Mises distribution and prior SNR," pp. 5063 – 5067, Apr. 2015.

[8] T. Gerkmann, M. Krawczyk-Becker, and J. L. Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 2, pp. 55–66, March 2015.

[9] F. Chen, L. L. Wong, and Y. Hu, "A Hilbert-fine-structure-derived physical metric for predicting the intelligibility of noise-distorted and noise-suppressed speech," *Speech Communication*, vol. 55, pp. 1011–1020, 2013.

[10] F. Chen and T. Guan, "Effect of temporal modulation rate on the intelligibility of phase-based speech," *J. Acoust. Soc. Am.*, vol. 134, pp. 520 – 526, 2013.

[11] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.

[12] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.

[13] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, "Phase-aware signal processing in speech communication: History, theory and practice," *John Wiley & Sons*, 2016.

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[15] T. Kleinschmidt, S. Sridharan, and M. Mason, "The use of phase in complex spectrum subtraction for robust speech recognition," *Elsevier Computer Speech and Language*, vol. 25, no. 3, pp. 585 – 600, 2011.

[16] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2015, pp. 708–712.

[17] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech r ecognition," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 1, pp. 190–202, Jan. 2007.

[18] S. H. Parthasarathi, R. Padmanabhan, and H. A. Murthy, "Robustness of group delay representations for noisy speech signals," *International Journal of Speech Technology*, vol. 14, no. 4, pp. 361–368, 2011.

[19] H. A. Murthy and B. Yegnanarayana, "Formant extraction from group delay function," *Speech Communication*, vol. 10, no. 3, pp. 209 – 221, 1991.

[20] Z. Tuske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," 2014, pp. 890–894.

[21] E. Loweimi, S. Ahadi, and T. Drugman, "A new phase-based feature representation for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2013, pp. 7155–7159.

[22] L. D. Alsteris and K. K. Paliwal, "ASR on speech reconstructed from short-time Fourier phase spectra," 2004, pp. 1–4.

[23] E. Loweimi, J. Barker, and T. Hain, "Source-filter separation of speech signal in the phase domain," 2015, pp. 598–602.

[24] R. Schlüter and H. Ney, "Using phase spectrum information for improved speech recognition performance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2001, pp. 133–136.

[25] P. Golik, Z. Tuske, R. Schlüter, and H. Ney, "Multilingual features based keyword search for very low-resource languages," 2015, pp. 26–30.

[26] J. Kulmer, P. Mowlaee, and M. Watanabe, "A probabilistic approach for phase estimation in single-channel speech enhancement using von Mises phase priors," in *IEEE Int. Workshop on Machine Learning for Signal Process.*, Sept. 2014.

[27] J. Kulmer and P. Mowlaee, "Phase estimation in single channel speech enhancement using phase decomposition," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 598–602, May. 2015.

[28] K. Mardia, *Statistics of Directional Data*. New York: Academic, 1972.

[29] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, Nov 2006.

[30] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[31] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 9, no. 5, pp. 504–512, 2001.

[32] M. Brookes, "VOICEBOX: Speech Processing Toolbox for MATLAB," Web page, 2005.

[33] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challange: An overview of challenge systems and outcomes," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 162 – 167, Dec 2013.

[34] H. G. Hirsch and D. Pearcs, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *ASR2000 - Automatic Speech Recognition: Challenges for the new Millenium*, pp. 181–188, Sep 2000.

[35] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, L. Xunying, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2006.

[36] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[37] B. Yegnanarayana and H. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Trans. Signal Process.*, vol. 40, no. 9, pp. 2281–2289, Sep 1992.