INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Speech Acoustic Modelling using Raw Source and Filter Components

Erfan Loweimi [1], *Zoran Cvetkovic* [2], *Peter Bell* [1] *and Steve Renals* [1]

[1] Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
[2] Department of Engineering, King's College London, UK

{e.loweimi, peter.bell, s.renals}@ed.ac.uk, zoran.cvetkovic@kcl.ac.uk

Abstract

Source-filter modelling is among the fundamental techniques in speech processing with a wide range of applications. In acoustic modelling, features such as MFCC and PLP which parametrise the filter component are widely employed. In this paper, we investigate the efficacy of building acoustic models from the raw filter and source components. The raw magnitude spectrum, as the primary information stream, is decomposed into the excitation and vocal tract information streams via cepstral liftering. Then, acoustic models are built via multi-head CNNs which, among others, allow for processing each individual stream via a sequence of bespoke transforms and fusing them at an optimal level of abstraction. We discuss the possible advantages of such information factorisation and recombination, investigate the dynamics of these models and explore the optimal fusion level. Furthermore, we illustrate the CNN's learned filters and provide some interpretation for the captured patterns. The proposed approach with optimal fusion scheme results in up to 14% and 7% relative WER reduction in WSJ and Aurora-4 tasks.

Index Terms: Source-filter separation, multi-stream information processing, raw spectrum-based acoustic modelling, ASR

1. Introduction

Source-filter modelling [1, 2] is among the fundamental techniques in speech processing. Based on the properties of human speech production system, this model characterises the speech signal as a temporal convolution of some random or quasi-periodic *excitation* (Exc) signal (*source*) passing through a linear *filter* representing the *vocal tract* (VT). These two elements can be separated via *deconvolution* in the time domain. *Cepstral low-pass liftering* (CLPL) [3,4] and *linear prediction* (LP) [5,6] are two popular methods for extracting the VT part.

It is well-established that the filter component is primarily associated with the linguistic content of the speech signal while the source element reflects the attributes correlated with the speaker [4]. Based on these properties, source-filter modelling has been widely applied in speech processing, e.g., in speech coding [7], speech synthesis [8–12], voice activity detection [13] and feature extraction for speech recognition/classification [14, 15]. Front-ends such as MFCC [14] and PLP [15], loosely speaking, parametrise the filter part using non-parametric CLPL and parametric LP methods, respectively.

In [16], we investigated the possibility of building acoustic models using different representations of the raw[1] phase spectrum including its source and filter components separated using the phase-based source-filter separation algorithm proposed in [17–19]. In this paper, we extend that study to the magnitude spectrum domain and scrutinise some unexplored aspects of acoustic modelling using raw source and filter components.

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

[1] By *raw* we mean using the entire spectrum (positive frequencies).

Having separated these two components via CLPL, we recombine these two information streams via multi-head convolutional neural networks (CNN) with multiple information fusion schemes. We also examine the dynamics of such approach in terms of evolution of the cross entropy (CE) loss and word error rate (WER) vs epoch. In addition, the learned filters of the first convolutional layer are depicted and some interpretations for the captured patterns are provided. The proposed framework could lead to up to 14% and 7% relative WER reduction in the WSJ [20] and Aurora-4 [21] tasks, respectively.

After briefly reviewing the source-filter modelling and separation in Section 2, we discuss the "why" and "how" of recombining the source and filter components via multi-head CNNs in Section 3. In Section 4 the learned filters are analysed and some interpretation for the captured patterns is presented. Section 5 is dedicated to the experimental results along with discussion and Section 6 concludes the paper.

2. Source-filter Separation

In this section, we briefly review the magnitude-based source-filter separation using CLPL [4]. This technique is a well-established method within the generic *Homomorphic Speech Analysis* [3] framework which aims at solving non-linear problems such as deconvolution via linear filtering in some domain.

The CLPL method is predicated on the assumption that the log of the magnitude spectrum (Fig. 1(b)) can be interpreted as a superposition of two components: a rapidly oscillating component modulated by a slowly varying element which are associated with the excitation and vocal tract parts, respectively. The slowly varying component (the vocal tract) can be extracted through convolution with a low-pass filter in the frequency domain or multiplication in a low-pass *lifter* in the *quefrency* (cepstral) domain. Fig. 1 illustrates the separated source and filer components via cepstral low-pass liftering.

To implement the low-pass liftering, we deploy a *brick-wall* (ideal low-pass) lifter of length L_0 in the quefrency domain which is equivalent to convolution with a cardinal sine (*sinc*) function in the frequency domain. To extract the VT part, L_0 should be slightly smaller than the fundamental periodicity, T_0. This requirement necessitates tracking T_0 per frame.

To keep the setup as simple as possible, we do not compute T_0 per frame and set the L_0 based on the minimum possible fundamental periodicity, T_0^{Min}. Assuming the maximum F_0 is 320 Hz (all speakers are adult) and the sampling rate is 16 kHz, we set L_0 to 50 (16000/320 samples). This ensures the filter component is devoid of any source information, but the excitation part is likely to include some VT residues.

It may be argued that by overlooking the actual T_0 per frame, the source-filter separation through CLPL accompanies with some notable error, namely the Exc part contains VT residues, and the larger the actual T_0, the higher the correspond-

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-53

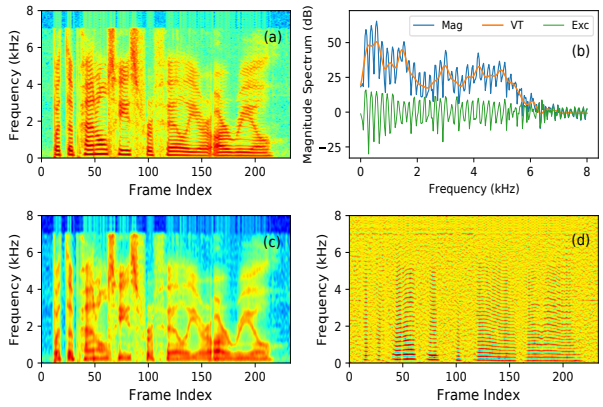

Figure 1: *Source-filter separation using cepstral low-pass liftering. (a) spectrogram, (b) Magnitude spectrum along with the VT and Exc components, (c) VT component, (d) Exc component.*

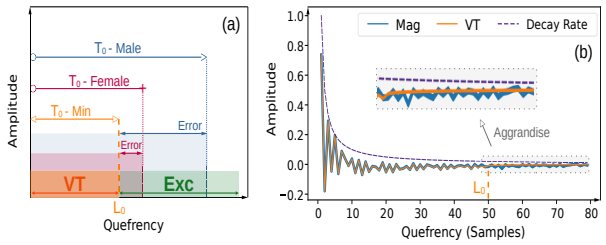

Figure 2: *(a) Error caused by using T_0^{Min} instead of the actual T_0 for a typical male/female speaker, (b) CCs' average tends to zero, decay rate is as fast as a Homographic function $(1/x)$.*

ing error (Fig. 2 (a)). However, considering one particular property of the cepstral coefficients (CC) proves that such error is practically negligible. In [3], it is shown that the CCs of the all-pole models (which well characterise the VT component) are inversely proportional with the quefrency index and the envelope decays quickly, as fast as a Homographic function $(1/x)$.

Fig. 2 (b) illustrates the average CCs calculated over 300 speech signals ($>$ 217k frames) of WSJ Eval-92 data. The CCs are computed using the the original and vocal tract magnitude spectra. As seen, the empirical results are consistent with the theory and the CCs for the quefrencies larger than 50 are almost zero. Therefore, the error due to using the T_0^{Min} instead of the actual T_0 per frame is insignificant. As such we can safely skip computing and tracking T_0 per frame which highly simplifies and facilitates applying the proposed approach.

3. Source-filter Recombination through Multi-head CNNs

Having separated the VT and Exc components, we wish to build an acoustic model using these two information streams without discarding any information. Our baseline system is a single-head CNN fed with the raw magnitude spectrum. Using the raw magnitude spectrum as input implies direct combination (multiplication or addition (log)) of the source and filter elements in the input (very low) level. We argue that deferring the combination to a higher level of abstraction, and pre-processing the information streams via multi-head CNNs before fusion, could lead to a more effective task-tailored information processing.

Now, we investigate the intuition and advantages of such framework as well as some implementation aspects.

3.1. Intuition and Advantages

What are the benefits of such factorisation and recombination? Information-wise, using the raw source and filter components simultaneously as inputs is as informative as using the raw magnitude spectrum. So, comparatively, such a multi-stream approach does not take advantage of any extra bits of information in the input level. As such the only remaining room to obtain a better performance is to process the supplied information in a more effective way. We believe this is the case and put forward three arguments lending support to this claim.

First, the *relevance* of each individual information stream to the given task is different. As shown in Tables 2 and 3, although the Exc component is obviously not as instrumental as the VT element in the ASR task, its performance is remarkably better than random guessing. This indicates that for a given task, the optimal weight for these two components before mixing them is not digital (0/1) but more subtle. We may also think about adequate weighting as a gating mechanism where ideally each stream should pass through a soft rather than a hard (0/1) gate. Using the magnitude spectrum along with a single-head CNN means giving these two components identical weights before fusion while we know a priori that their importance is different.

Second, regardless of the importance of each information stream to the task, the underlying information generation process encodes the information within different *forms* and/or set of patterns. As such optimal chain of transforms for processing and distilling the carried information by each stream would be dissimilar. For example, for tonal languages, the source component includes some information correlated with the lingual content of speech. However, the optimal pipeline for processing this information stream is entirely different from the optimal pipeline for processing the vocal tract stream. The proposed framework can handle this issue by learning a set of bespoke transforms which process each stream individually and then fusing them in an optimal level of abstraction.

Third, an optimal information processing system ideally should only pass through the task-relevant information to the decision making level whilst filters out the irrelevant components. If the model is fed with multiple streams, it can not only learn what is more important to the task and pass it through, but also learn which data variability is irrelevant and hence should be neglected during inference. In ASR, such information filtering, among others, means removing the speaker-related attributes. The raw source component is a simple yet rich representation encoding such attributes and feeding the system with it helps the model to learn to normalise/ignore such irrelevant data variability. This potentially contributes towards building a system with a better generalisation and robustness.

3.2. Implementation

Having processed each individual information stream, we shall fuse them at some point along the pipeline. Fig. 3 illustrates three multi-stream architectures along with the single-head baseline system. As seen, the systems composed of a cascade of the convolutional and fully-connected (FC) sub-networks, fusing the information streams at three levels: low level after the convolutional sub-network (**Concat-1**), medium level in the middle of the FC sub-network (**Concat-2**), and high level, just before the output (softmax) layer (**Concat-3**).

The processed information streams should be mixed when they reach an optimal level of abstraction, which is determined empirically, depending on the architecture, streams' information content, data and task. However, the following could

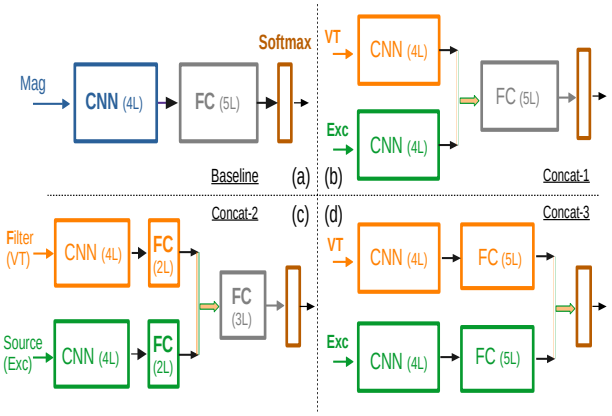

Figure 3: *Multi-head acoustic modelling using convolutional and fully-connected (FC) sub-networks (n in nL is number of layers). Fusion (concatenation) of the raw magnitude spectra of the source and filter components at different levels: (a) Baseline (single-head), (b) Concat-1, (c) Concat-2, (d) Concat-3.*

Table 1: *#params (in millions) for systems in Fig. 3.*

	Baseline	Concat-1	Concat-2	Concat-3
#params	9.6	12.9	15.1	19.3

elucidate some of the advantages and drawbacks of different schemes: i) assuming there is a fixed budget in terms of number of layers, placing the fusion point at the higher levels leads to allocating more layers to pre-process individual streams, leaving fewer layers and consequently capacity for post-processing and abstraction extraction after the fusion point; ii) the higher the fusion level, the higher the number of architecture parameters (*#params*), as shown in Table 1.

4. Interpretation of the Learned Filters

In this section, we depict the learned filters in the first convolutional layer (ConvL1) for the respective models and provide some interpretation for the captured patterns. Filters referred to as *Mag* in Figs. 4-6 belong to the single-head baseline system (Fig. 3 (a)) fed with the original magnitude spectrum. Those referred to as *VT* and *Exc* are related to the Concat-1 architecture (Fig. 3 (b)) fed with the raw magnitude spectra of the VT and Exc components, respectively. Each ConvL1 has 128 filters and the kernel size is 129. In all of the experiments, 10^{th} root of the magnitude spectrum is used as input. The task is WSJ.

Fig. 4 illustrates the learned filters (left column) along with the corresponding Fourier transform (FT) (right column). Although the filters are learned in the frequency domain, (in the left column of Fig. 4) we did not label the domain as frequency. However, we refer to the domain after taking FT of the filters as quefrency. One interesting observation is that the learned filters in the model fed with the the original raw magnitude spectrum emphasise mostly the low cepstral components (Fig. 4 (b)), *i.e.* the model identifies and focuses on the most important aspect of the input to the task, namely the part associated with the VT.

Another noteworthy observation is that the FT of the learned filters fed with VT (Fig. 4 (d)) is almost zero for quefrencies larger than 50. This illustrates that the model, among others, learns about an important property of its input, namely the magnitude spectrum of the vocal tract component, and does not pay attention to higher quefrencies which are already set to

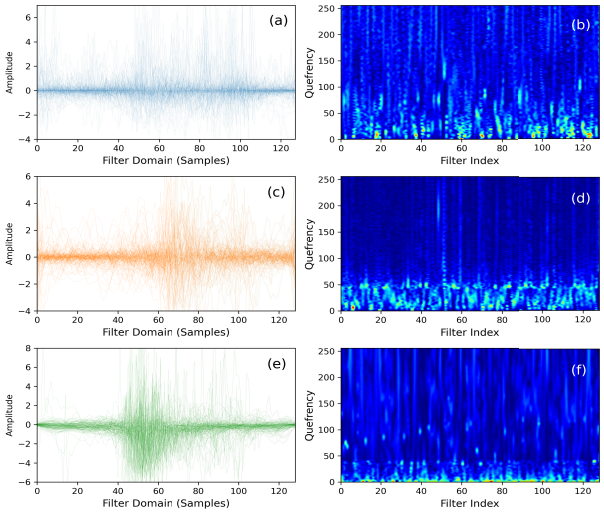

Figure 4: *Learned filters in the first convolutional layer fed with Mag, VT and Exc components; left column is filters directly operate on the magnitude spectra while right column depicts their FT (quefrency). (a) and (b) Mag, (c) and (d) VT, (e) and (f) Exc.*

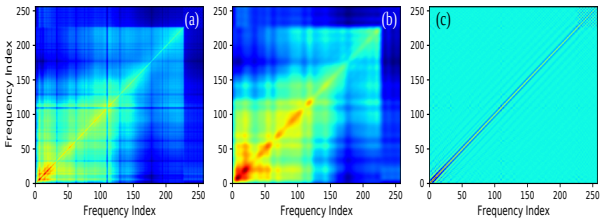

Figure 5: *Covariance matrices computed using 300 signals of WSJ Eval-92. (a) Magnitude spectrum, (b) VT, (C) Exc.*

zero during VT extraction via CLPL. This property is comparable with raw waveform models which can detect noisy subbands and filter them out, as demonstrated in Fig. 1 (h) in [22].

Another dimension to explore is the support of the learned filters, which might be understood by comparing with the covariance matrices of the raw Mag, VT and Exc components. The filters should capture some task-useful local correlations in the input. Based on the empirically computed covariance matrices using WSJ Eval-92 data (Fig. 5), short, mid and long-range dependencies/correlations are all plausible, especially for Mag and VT. This justifies having filters with short to long supports.

The last insight regarding the learned filters is about their shape. As seen in Fig. 6, the filters could be unimodal, resembling a wavelet such as a sinc function with a short (Fig. 6 (a)) or medium support (Fig. 6 (b)), or bimodal with medium to long support (Fig. 6 (c)), or could have a very long support (Fig. 6 (d)). Similarity with the sinc parametric filters also implies that as well as conventional non-parametric filters, one may use parametric CNNs [23] as the first layer, e.g., SincNet [24]. We take advantage of this observation in the next section.

5. Experimental Results

5.1. Setup

DNNs were trained using PyTorch-Kaldi [25, 26] with default recipes; raw waveform model configuration has been used for processing the raw spectra. The baseline architectures (Fig. 3

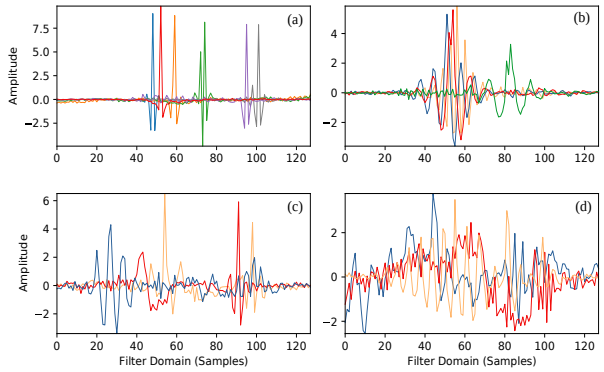

Figure 6: *Shapes of the learned filters in ConvL1 fed with the raw magnitude spectrum (Fig. 4 (a)). (a) Sinc (short), (b) Sinc (medium support), (c) bimodal, (d) long-range correlation.*

(a)) consists of a cascade of four 1D convolutional layers followed by five FC hidden layers with ReLU [27] activation. Experiments where carried out on Aurora-4 (multi-style) [21] and WSJ [20] tasks. Alignments were taken from the respective Kaldi standard recipes [28]. For both tasks the WER and CE loss are reported. Aurora-4 test set includes four subsets: A (clean), B (additive noise), C (channel noise) and D (additive and channel noise). Length of the MFCC, FBank and raw features (per frame) are 39, 40 and 257, respectively. $\text{Mag}^{0.1}$ indicates 10^{th} root. VT and Exc in Tables 2 and 3 are 10^{th} root of the corresponding magnitude spectra. Feature vectors were augmented with the features of the ± 5 contextual frames.

5.2. Results and Discussion

Tables 2 and 3 show the WER for WSJ and Aurora-4, respectively. As seen, presenting the network with the source and filter components instead of the magnitude spectrum, whilst not providing any new information, leads to a notable performance gain owing to a more effective information processing.

Information fusion level plays a noticeable role. While Concat-3 leads to the poorest performance, Concat-1 appears to be the optimal fusion scheme resulting in up to 14% and 5% relative (to the raw magnitude spectrum with single-head) WER reduction for WSJ and Aurora-4 tasks, respectively.

Fig. 7 illustrates the temporal evolution (dynamics) of the CE and WER vs epoch for different systems and datasets. As seen, for WSJ the performance measures reach a plateau after 15 epochs while for Aurora-4 the performance keeps improving for up to 30 epochs. Such differences in dynamics and performance gain are explainable considering the amount of training data (81h vs 14h) and presence of noise in Aurora-4.

Motivated by the captured patterns depicted in Fig. 6 (a) and (b) where some filters resemble parametric models, we also studied the usefulness of employing CNNs with parametric kernels and particularly, SincNet. Such models were primarily proposed for acoustic modelling from the raw waveform. Here, we successfully extended their use-case to the raw spectrum-based acoustic modelling. As seen in Tables 2 and 3, this modification further improves the results, leading to WER of 4.5% and 8.1% for WSJ (Eval-92) and Aurora-4 (average), respectively.

6. Conclusions

In this paper, we investigated the usefulness of acoustic modelling for ASR using raw magnitude spectra of the source and

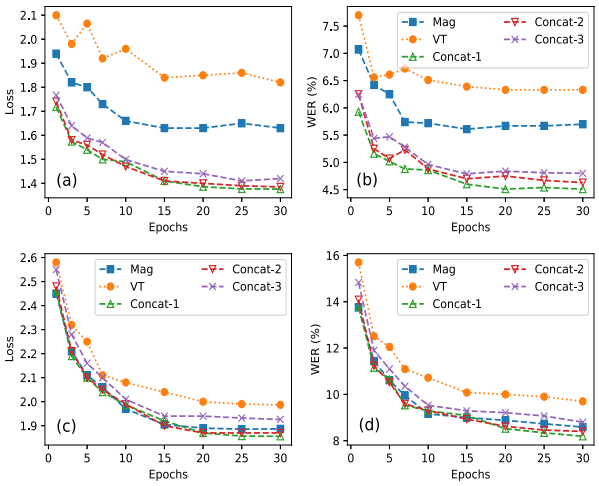

Figure 7: *Temporal evolution (dynamics) of the CE and WER vs epoch. (a) CE for WSJ (Dev-93), (b) WER for WSJ (Eval-92), (c) CE for Aurora-4 (Dev), (d) Average WER for Aurora-4.*

Table 2: *WSJ WER for different front-ends.*

	Dev	Eval-92	Eval-93
MFCC	10.4	6.8	10.4
FBank	9.1	5.9	8.8
Raw-wave	8.7	5.2	8.5
$\text{Mag}^{0.1}$ (baseline)	8.8	5.5	9.0
Exc	15.1	11.8	16.5
VT	9.6	6.3	9.1
Concat-1	7.9	4.5	7.5
Concat-2	7.9	4.6	7.6
Concat-3	8.1	4.8	7.6
Sinc-Concat-1	8.0	4.5	7.4

Table 3: *Aurora-4 (multi-style) WER for different front-ends.*

Feature	A	B	C	D	*Avg*
MFCC	3.5	6.8	7.1	16.5	10.7
FBank	2.9	5.9	4.5	14.5	9.2
Raw-wave	3.1	5.7	7.5	16.5	10.3
$\text{Mag}^{0.1}$ (baseline)	2.6	5.3	4.3	14.1	8.8
VT	3.0	6.0	5.1	15.0	9.6
Exc	6.4	15.8	16.2	32.6	22.4
Concat-1	2.4	5.1	4.1	13.0	8.2
Concat-2	2.5	5.2	4.3	13.3	8.4
Concat-3	2.5	5.5	4.5	13.9	8.8
Sinc-Concat-1	2.3	5.0	4.0	12.7	8.1

filter components. Having separated the vocal tract and excitation elements of the speech signal, these two streams of information were recombined via multi-head CNNs. Advantages of such factorisation and recombination were discussed. It was argued that it paves the way for a more effective task-oriented multi-stream information processing. Performance-wise, up to 7% and 14% relative WER reduction for Aurora-4 and WSJ tasks have been achieved. Training dynamics and optimal fusion schemes were explored, the learned filters were analysed and some interpretation for the captured patterns were presented. The proposed multi-stream source-filter-based approach provides a generic framework, potentially employable in a wide range of speech recognition and classification tasks.

7. References

[1] T. Chiba and M. Kajiyama, *The vowel, its nature and structure*. Phonetic Society of Japan, 1958.

[2] G. Fant, *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*, ser. Description and Analysis of Contemporary Standard Russian. De Gruyter, 1971.

[3] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, pp. 221–226, Jun 1968.

[4] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[5] B. S. Atal and S. L. Hanauer, "peech analysis and synthesis by linear prediction of the speechwave," *The Journal of the Acoustical Society of America*, vol. 50, pp. 637–655, 1971.

[6] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[7] M. Schroeder and B. Atal, "Code-excited linear prediction(celp): High-quality speech at very low bit rates," in *ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 10, Apr 1985, pp. 937–940.

[8] J. L. Flanagan, *Speech Analysis; Synthesis and Perception*. Springer-Verlag, 1972.

[9] J. L. Baart and V. J. van Heuven, "From text to speech; the mitalk system: Jonathan allen, m. sharon hunnicutt and dennis klatt (with robert c. armstrong and david pisoni): Cambridge university press, cambridge, 1987. xii+216 pp. £25.00," *Lingua*, vol. 81, no. 2, pp. 265–270, 1990.

[10] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[11] J. P. Cabral, K. Richmond, J. Yamagishi, and S. Renals, "Glottal spectral separation for speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 195–208, 2014.

[12] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 402–415, 2020.

[13] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice activity detection: Merging source and filter-based information," *IEEE Signal Processing Letters*, vol. 23, no. 2, pp. 252–256, 2016.

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 357–366, 1980.

[15] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

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

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

[18] E. Loweimi, J. Barker, O. Saz Torralba, and T. Hain, "Robust source-filter separation of speech signal in the phase domain," in *INTERSPEECH*, 2017, pp. 414–418.

[19] E. Loweimi, "Robust phase-based speech signal processing; from source-filter separation to model-based robust asr," Ph.D. dissertation, University of Sheffield, 2018. [Online]. Available: http://etheses.whiterose.ac.uk/19409/

[20] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *IEEE ICASSP*, 1992, pp. 899–902.

[21] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," Inst. for Signal and Information Process, Mississippi State University, Tech. Rep., 2002.

[22] E. Loweimi, P. Bell, and S. Renals, "On the Robustness and Training Dynamics of Raw Waveform Models," in *INTERSPEECH*, 2020, pp. 1001–1005.

[23] ——, "On learning interpretable CNNs with parametric modulated kernel-based filters," in *INTERSPEECH*, 2019.

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

[25] M. Ravanelli, T. Parcollet, and Y. Bengio, "The PyTorch-Kaldi speech recognition toolkit," in *IEEE ICASSP*, 2019.

[26] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Workshop on Autodiff*, 2017.

[27] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *AISTATS*, 2011, pp. 315–323.

[28] 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 ASRU*, 2011.