A Study of Salient Modulation Domain Features for Speaker Identification

Simon W. McKnight, Aidan O. T. Hogg, Vincent W. Neo, Patrick A. Naylor
Department of Electrical and Electronic Engineering
Imperial College London, UK
{s.mcknight18, aidan.hogg13, vincent.neo09, p.naylor}@imperial.ac.uk

***Abstract*—This paper studies the ranges of acoustic and modulation frequencies of speech most relevant for identifying speakers and compares the speaker-specific information present in the temporal envelope against that present in the temporal fine structure. This study uses correlation and feature importance measures, random forest and convolutional neural network models, and reconstructed speech signals with specific acoustic and/or modulation frequencies removed to identify the salient points. It is shown that the range of modulation frequencies associated with the fundamental frequency is more important than the 1-16 Hz range most commonly used in automatic speech recognition, and that the 0 Hz modulation frequency band contains significant speaker information. It is also shown that the temporal envelope is more discriminative among speakers than the temporal fine structure, but that the temporal fine structure still contains useful additional information for speaker identification. This research aims to provide a timely addition to the literature by identifying specific aspects of speech relevant for speaker identification that could be used to enhance the discriminant capabilities of machine learning models.**

I. Introduction

Speaker identification remains one of the most important unsolved problems that has many crucial applications. Many methods have been proposed and tested for generating suitable features from speech that can be used to identify speakers, including hearing perception tests in cognitive psychology and physical models of human voice production. In recent years, machine learning models have come to dominate speaker identification research, and end-to-end systems are particularly popular. However, a major drawback of those systems is that they are effectively black box models, meaning that although they produce state-of-the-art results, they provide little understanding about which specific components of speech are important for generating specific speaker identifier outputs as well as requiring a substantial amount of training.

This paper studies features generated from the modulation spectrum and their application to speaker identification to bridge the gap. This allows specific acoustic frequencies and modulation frequencies to be tested and their effects on speaker identification systems investigated.

A. Speaker Identification Background

Speaker recognition refers to the broad category of systems that use voice features generated from speech to identify the speaker [1], [2]. There are three main categories of speaker recognition: (a) speaker identification, which is a 1-of-N problem in that the system is trained on N specific speakers and used to identify which speaker in the training set a particular test recording belongs (closed-set), if any (open-set); (b) speaker verification, which is a 1-to-1 problem in that the system is trained to assess whether a particular test recording belongs to one particular speaker or not; and (c) speaker diarization, which is where speakers in a particular test recording are distinguished and the times at which each speaker was speaking is identified. Speaker diarization differs from speaker identification and speaker verification in that (i) the issue of overlapping speakers (i.e. more than one person speaking at the same time) is more significant and (ii) it is usually applied on an unsupervised basis in that the system distinguishes speakers that it may not have been trained on.

All areas of speaker recognition are active areas for research and regular challenges are held to encourage new research in particular directions, including: (i) the National Institute of Standards and Technology (NIST) speaker recognition evaluation challenges on speaker identification and speaker verification with certain challenge-specific variations [3], [4], (ii) the VoxCeleb speaker recognition challenges on speaker verification and speaker diarization [5]; and (iii) the DIHARD I, II and III challenges [6] on speaker diarization.

The most commonly used single frame features for speaker identification are mel-frequency cepstral coefficients (MFCCs), though mel filter bank cepstral coefficients (FBANKs), linear predictive cepstral coefficients (LPCCs) and coefficients derived using perceptual linear prediction (PLP) are also popular [7]. These single frame features are hand-crafted and generated in an unsupervised manner. They describe both phonetic and speaker characteristics, but the former should ideally be avoided or ignored in speaker identification.

It has been known for some time that using features generated across a number of frames contains useful speaker information that is not evident in single frame based features alone [8], [9]. Many methods have been proposed and tested for generating segment-level features based on multiple frames, ranging from simple delta and delta-delta features (also called velocity and acceleration) that show the rates of change across two or more frames to more sophisticated models. Unsupervised methods have the advantage of not needing substantial training, but in practice best results are currently obtained using supervised methods. State-of-the-art speaker identification performance is currently obtained using,

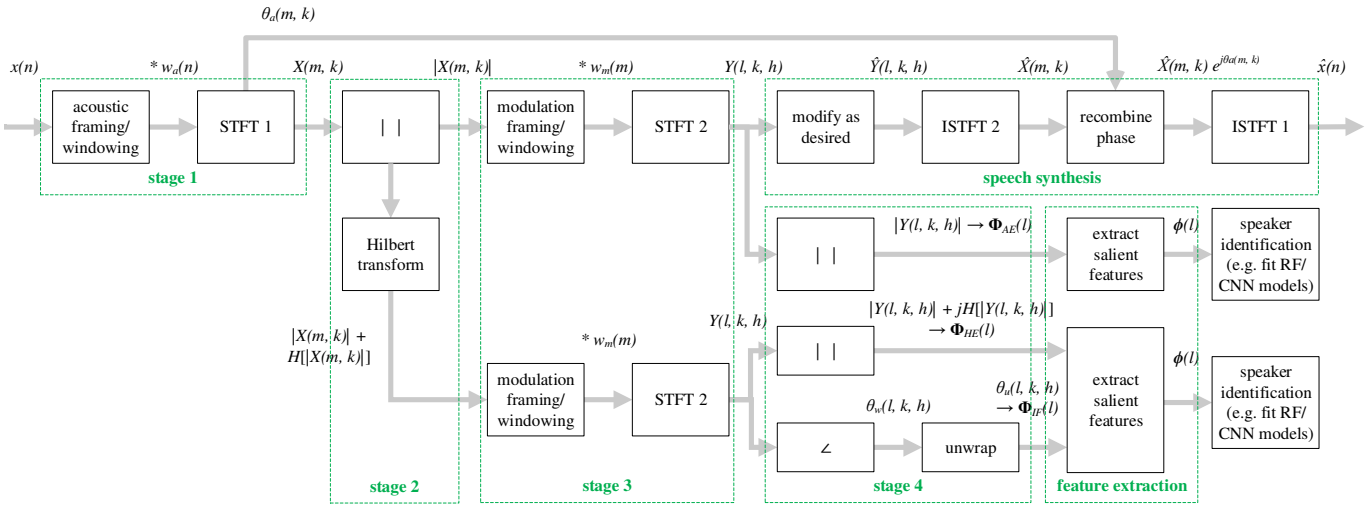

Fig. 1: Modulation spectrum generation flowchart.

for example, an extended time delay neural network (TDNN) based on MFCCs [10], [11] or FBANKs [12] to generate speaker embeddings, x-vectors, from the penultimate layer of the TDNN. Other multi-frame features include modulation spectrum features described in Section I-B, which are the primary focus of this paper.

The other important feature relevant for this paper that is sometimes used in conjunction with MFCCs or FBANKs is the fundamental frequency (F0) expressed in Hz, often referred to as the pitch [13] (there are other distinctions between F0 and pitch that are not addressed here, but this paper will refer to F0 only going forward). F0 varies in a range for individual speakers, around 85-155 Hz for male speakers and 165-255 Hz for female speakers [14], and different speakers with similar F0 can sound completely different so it is insufficient on its own for speaker identification. Much of the research on F0 focuses on the shape of the F0 contour over a range of frames (e.g. F0 of a speaker should change smoothly during an utterance, so significant jumps could indicate a different speaker). One drawback of F0 is that it is only available for voiced speech, not unvoiced. Also, MFCCs and FBANKs are known to have some intrinsic F0 information [15]. Although this paper does not calculate F0 directly, it is relevant for, and directly comparable to, the analysis of the fundamental frequencies in the modulation domain in this paper.

B. Modulation Spectrum Background

The modulation spectrum describes how the acoustic frequency components of a speech signal change over time [16]. There are two parts: (i) the temporal envelope, which uses amplitude modulation (AM) principles to look at the slowly changing temporal trajectory of specific acoustic frequency bands in the speech spectrum; and (ii) the temporal fine structure, which uses frequency modulation (FM) principles to look at the rapidly changing instantaneous frequency around the centre frequencies of those same acoustic frequency bands in the speech spectrum. Much of the existing research into the modulation spectrum of speech focuses on the temporal envelope as human hearing has generally been shown to be more sensitive to AM than to FM or phase modulation (PM), and that most linguistic information is contained in specific parts of the temporal envelope [17]. However, a growing body of research shows that FM features are helpful for distinguishing speakers, as discussed in Section I-B2 below.

The steps used to generate the modulation spectrum features are described in Section II and the process flowchart is shown in Fig. 1. References to "modulation frequency" without any other qualifier mean the frequency of the temporal envelope.

1) Temporal Envelope Literature: Early research into the intelligibility of Dutch speech when manipulating specific parts of the modulation spectrum found that there was no effect on speech intelligibility when (a) modulation frequencies above 16 Hz were removed, provided all lower modulation frequencies were present [18], and (b) modulation frequencies below 4 Hz were removed while retaining all higher modulation frequencies [19]. This does not mean only modulation frequencies in the range 4-16 Hz are relevant for speech intelligibility though, and subsequent research on English and Japanese speech has identified the 1-16 Hz range as containing most useful linguistic information about speech for automatic speech recognition (ASR) with the 2-8 Hz range particularly important and peaking at 4 Hz [20], [21], [22]. Several studies investigated specific aspects of speech at specific modulation frequencies, which [23] summarises as broadly: stress rate 1-2 Hz; syllable rate 2-8 Hz; and phoneme rate 8-40 Hz. However, [23] notes there is some variability in these rates across studies. Humans were shown to understand temporal envelope frequencies up to 256 Hz in [24]. Other research looked at temporal envelope frequencies up to 250 Hz or more so the fundamental frequencies of adult speakers was visible in the modulation domain [25], [26].

2) Temporal Fine Structure Literature: The importance of AM and FM on speech recognition and speaker identification was studied in [27] in both English and Mandarin, finding that AM was sufficient for speech recognition in quiet environments but that the addition of FM substantially improved speech recognition in noisy environments, speaker identification and tonal language recognition. Furthermore,

FM was found to be critically important in distinguishing speech from overlapping speakers, which the paper referred to as a "competing voice". The test subjects comprised both people with normal hearing and people with cochlear implants. Similarly, [28] found that different instantaneous frequencies within speech were significant for identifying the speaker, and developed features based on the AM-FM representation of speech that were shown to be robust to changes in the recording channel and speaking style. Subsequently, [29] developed an AM-FM filter bank that converted speech to a spectro-temporal representation with less smearing or scattering, and consequently better speaker identification performance, than a typical discrete cosine transform (DCT) form. Another helpful analysis comes from [30], which found that the temporal fine structure aided speech segmentation, though that analysis was solely based on Mandarin speakers.

In [31], the importance of temporal fine structure on masking, pitch perception and speech perception was investigated, finding that it is important for all three. That research built on an earlier paper by the same author that analysed the then-prevailing view that the human ear is insensitive to phase [32]. The concept of instantaneous frequency cosine coefficients (IFCCs) was developed in [33], and was shown to improve speaker verification performance when used in conjunction with MFCCs and frequency domain linear prediction (FDLP) features. Further research on the temporal fine structure (e.g. [34], [35]) supports the importance of retaining information derived from it to obtain all the cues inherent in speech, which is a particular problem in cochlear implants.

3) Modulation Spectrum in Speaker Recognition: Previous work on the application of the modulation spectrum to speaker recognition started with [36] and analysed properties of the modulation spectrum relevant for speaker verification in continuous telephone speech sampled at 8 kHz using various modulation filters on acoustic feature bands, finding that spectral components between 0.1 and 10 Hz have the most useful speaker information, with frequencies below 0.125 Hz in both matched and mismatched conditions and above 8 Hz in mismatched conditions even being detrimental to speaker recognition (in that research, matched conditions means that the telephone handset type was the same as the telephone number type based on NIST labels). The 0.1 to 10 Hz range is particularly significant as it it closely matched the bandpass filter range used to convert PLP coefficients to relative spectral perceptual linear prediction (RASTA-PLP) coefficients after the speech signal was first broken up into critical bands [37]. However, both [36] and [37] were limited to applying the modulation filters to all acoustic feature bands, so [25] developed a joint acoustic-modulation frequency representation of speech (i.e. the modulation spectrum) that could identify important modulation frequencies at specific acoustic frequencies, and which highlighted that there were distinct groupings for two overlapping speakers. This idea was further developed in [38], before the same author investigated dimension reduction of the modulation spectrum for speaker recognition in [39]. More recently, [40] investigated the modulation spectrum using a discriminability index, finding that it was important for assessing "speaker individuality" and "vocal-emotion recognition".

In speaker diarization, [41] applied modulation spectrogram (MSG) features in conjunction with MFCCs, finding that together they significantly improved results. The DiarTk speaker diarization toolkit uses the modulation spectrum with MFCCs, time difference of arrival (TDOA) and FDLP features to generate their best results [42], [43], [44]. The modulation spectrum features are briefly described in [42] as "slowly varying components", which clearly refers to the temporal envelope, and "critical band energy trajectories are filtered using a low pass filter and the resulting features are decorrelated", but insufficient information is given about how the modulation spectrum features are generated.

C. This Research

This research is a study of modulation spectrum features that are most relevant for text-independent closed-set speaker identification. Section II discusses how the modulation spectrum features are generated and the variations available. Section III analyses how well those modulation spectrum features identify speakers using the datasets and systems described in Section III-A. It starts by analysing the feature correlations in Section III-B before quantifying how well modulation spectrum features identify speakers when used in supervised machine learning models in Sections III-C and III-D. Section III-E discusses modification and reconstruction of the speech signal from parts of the modulation spectrum as that highlights audible contributions made in parts of the modulation spectrum. The literature reviews in Section I investigated applications of the modulation spectrum to speaker verification and speaker diarization as well as to speaker identification as there is considerable overlap in the features and methods used.

The original contributions of this paper show (a) that the range of modulation frequencies associated with the fundamental frequency, rather than the 1-16 Hz range most commonly used in ASR, are particularly important for speaker identification, (b) that the modulation frequency band around 0 Hz modulation frequency contains significant speaker information and (c) that although the temporal envelope is more discriminative among speakers than the temporal fine structure, the temporal fine structure still contains useful additional information for speaker identification.

II. Modulation Spectrum Analysis

This paper employs a modified version of the method used in [45], omitting the two-dimensional discrete cosine transform (2D-DCT), to obtain the temporal envelope modulation spectrum features $\mathbf{\Phi} \in \mathbb{R}^{L \times K \times H}$, where L is the number of modulation frames, K is the number of acoustic frequency bands and H is the number of modulation frequency bands. In Sections III-C and III-E, a comparison is made between $\mathbf{\Phi}$ calculated using the amplitude envelope $\mathbf{\Phi}_{AE}$ in both stages 2 and 4 below and that calculated using the Hilbert envelope $\mathbf{\Phi}_{HE}$ in those stages. This paper also calculates the instantaneous frequencies (i.e. based on the temporal fine

structure) as a 3D tensor $\mathbf{\Phi}_{IF} \in \mathbb{R}^{(L-1)\times K\times H}$, except that for the purpose of Section III-C a different formulation is used for the final model that results in $\mathbf{\Phi}_{IF} \in \mathbb{R}^{L\times K\times H}$.

The process of generating $\mathbf{\Phi}$ and $\mathbf{\Phi}_{IF}$ from the modulation spectrum is described as a 4-stage process in this paper, with an additional last stage to describe options not addressed in the 3-stage process of [16]. Two of the stages involve taking a short-time discrete Fourier transform (STFT) [46], though other ways of producing frequency bands using some form of filter (either perceptually motivated or otherwise) would be equally valid [47]. The process flowchart is shown in Fig. 1.

The first stage uses the STFT to convert the acoustic speech signal $x(n)$ into a matrix $\mathbf{X} \in \mathbb{C}^{M\times K}$ with M acoustic frames and K acoustic frequency bands. Each element of $\mathbf{X}$ is

$$X(m,k) \;=\; \frac{1}{A}\sum_{i=0}^{W_a f_s-1} x(mF_a f_s \,+\, i)w_a(i)e^{-j\frac{2\pi ki}{W_a f_s}} \quad (1)$$

for acoustic frame index m and acoustic frequency band k, where i is the iterator over each sampling point in the window, f_s is the sampling frequency, W_a is the acoustic frame duration in seconds, F_a is the acoustic frame step in seconds, $w_a(i)$ is the acoustic window function, $\{m \in \mathbb{Z} : 0 \le m \le M-1\}$, $\{k \in \mathbb{Z} : 0 \le k \le K\}$, $K = \frac{W_a f_s}{2}$ after applying the Nyquist cut-off frequency in the acoustic domain, and the acoustic window scaling factor $A \;=\; \sum_{i=0}^{W_a f_s-1} w_a(i)$.

It is common to apply mel filter banks to the first stage STFT, and then sometimes to apply a DCT, resulting in FBANKs and MFCCs respectively. However, the aim of this paper is to pinpoint the salient features, so higher resolution frequency bands are used here.

The second stage finds the spectral envelope of specific frequency bands. One approach is to obtain $\mathbf{Z} \in \mathbb{R}^{M\times K}$ where each element is given by $|X(m,k)|$ and $|\cdot|$ denotes the modulus operator. Another approach is to use the discrete Hilbert transform denoted $H[\cdot]$, in which each element of $\mathbf{Z}$ is

$$Z(m,k) = \big|X(m,k)\big| + jH\Big[\big|X(m,k)\big|\Big]. \quad (2)$$

The phase of $Z(m,k)$ is denoted $\theta_w(m,k)$ for the wrapped instantaneous phase. That is converted to the unwrapped phase $\theta_u(m,k)$, so that the instantaneous frequency is then calculated as

$$f_i(m,k) = \frac{1}{2\pi}\left[\frac{\theta_u(m+1,k) - \theta_u(m,k)}{F_a}\right]. \quad (3)$$

The third stage applies a second STFT to a certain number of acoustic frames $X(m,k)$ in the same acoustic frequency band k and generates a tensor $\mathbf{Y} \in \mathbb{C}^{L\times K\times H}$. Each element

$$Y(l,k,h) \;=\; \frac{1}{B}\sum_{i=0}^{\frac{W_m}{F_a}-1} Z(l\frac{F_m}{F_a} + i, k)w_m(i)e^{-j\frac{2\pi F_a hi}{W_m}} \quad (4)$$

for modulation frame index l, acoustic frequency band k and modulation frequency band h, where i is the acoustic frame index within the modulation frame, W_m is the modulation frame duration in seconds (assumed to be an integral multiple of F_a, as is F_m), F_m is the modulation frame step in seconds, $w_m(i)$ is the modulation window function, $\{l \in \mathbb{Z} : 0 \le l \le L-1\}$, $\{k \in \mathbb{Z} : 0 \le k \le K\}$ as before, $\{h \in \mathbb{Z} : 0 \le h \le H\}$, $H = \frac{W_m}{2F_a}$ after applying the Nyquist cut-off frequency in the modulation domain and the modulation window scaling factor $B = \sum_{i=0}^{\frac{W_m}{F_a}-1} w_m(i)$.

In stage 4, the modulation spectrum $\mathbf{\Phi} \in \mathbb{R}^{L\times K\times H}$ is obtained as either $|Y(l,k,h)|$ for each element or by applying the Hilbert transform to find the Hilbert envelope. For the latter, the instantaneous frequencies $\mathbf{\Phi}_{IF} \in \mathbb{R}^{(L-1)\times K\times H}$ are also calculated. The methodology is the same as for stage 2.

A drawback of $\mathbf{\Phi}$ on its own is that for each modulation frame l, $\mathbf{\Phi}(l) \in \mathbb{R}^{K\times H}$ is a 2D matrix and usually needs to be converted into a low-dimension vector $\boldsymbol{\phi}(l) \in \mathbb{R}^{D\times 1}$ for use in speaker identification systems, where D denotes the reduced number of dimensions per modulation frame ($D = K \times H$ if the 2D matrix $\mathbf{\Phi}(l)$ is simply flattened to create $\boldsymbol{\phi}(l)$). Deep neural networks such as a convolutional neural network (CNN) are a popular choice as the speaker embeddings vector from the penultimate layer can be used. Discriminative features can then be made generative, e.g. using probabilistic linear discriminant analysis (PLDA) [48].

III. Experimental Design and Results

A. Datasets and Systems Used

A rearranged version of the TIMIT dataset [49] was employed. Although TIMIT was designed for ASR research, the data is conveniently arranged by speaker and utterance. Although the dataset is quite small, it provides a useful starting point for this research and enabled machine learning models to be trained relatively quickly. Future developments of this research will use larger datasets commonly used in modern speaker recognition challenges (e.g. the VoxCeleb 1 and 2 datasets [5]). To use TIMIT in speaker identification, the data was rearranged so that the training set comprises the first 7 utterances of each speaker (including SA1 and SA2) and the test set the last 3 utterances of each speaker.

Models were obtained to identify speakers using random forests [50] as, unlike some machine learning methods, they are readily interpretable and generate quantifiable feature importances. Using 100 constituent trees was found to be a good compromise between reducing overfitting and memory/storage limitations. The Gini impurity measure [50] was used with random selection with replacement. No leaf maximum depth or pruning was applied.

CNN models have in recent years produced good results in image recognition. Since the modulation spectrum comprises a 2D image per modulation frame, it is natural to fit a supervised CNN model on the training set and test how well it makes predictions on the test set. The CNN structure shown in Table I has also been investigated. Good experiment results were obtained training for 100 epochs using 10% spatial dropout after each convolution layer, 30% dropout after each dense layer and "same" filters with stride 1 for convolutions. A random 15% of the training data was allocated to a validation set to monitor accuracy and loss improving as expected. This CNN structure has over 7 million weights. While the TIMIT dataset is adequate for identifying salient modulation spectrum

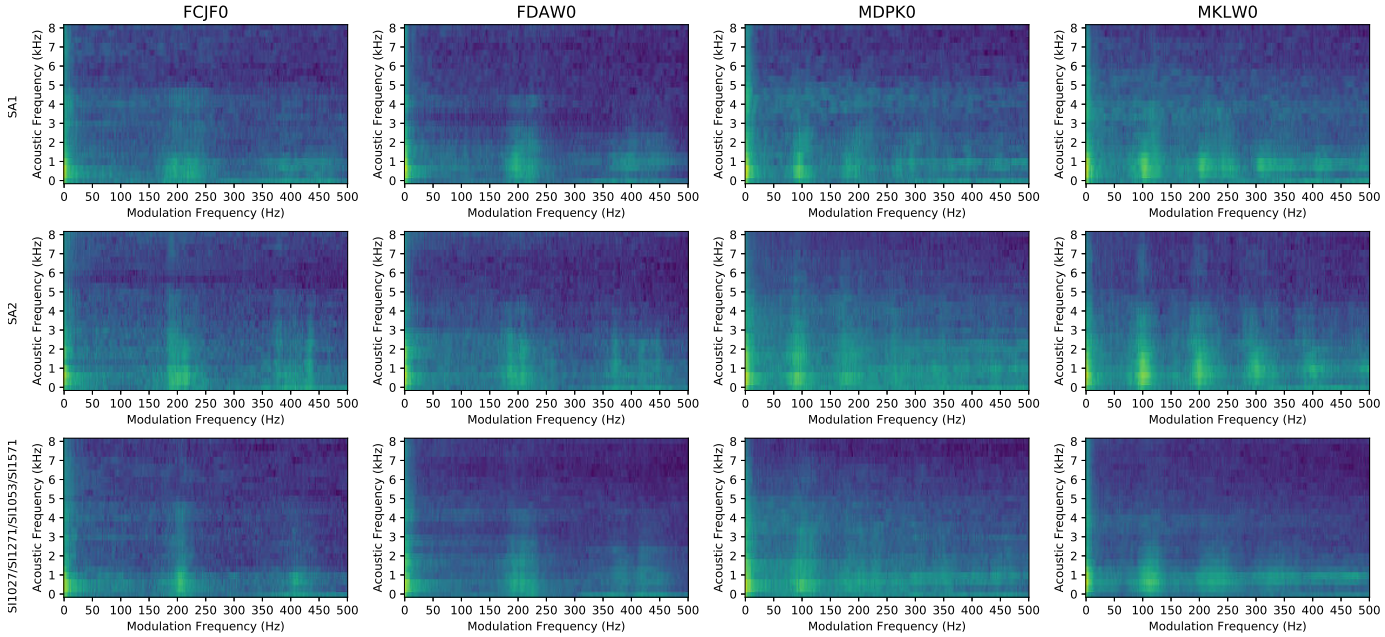

Fig. 2: 4 TIMIT speakers (2 female, 2 male) average modulation spectra by utterance shown in each row.

features, the CNN models are undertrained and results would be expected to improve using larger datasets.

The wideband modulation spectrum used throughout this research was generated using $F_a = 1$ ms, $W_a = 3$ ms, $F_m = 100$ ms and $W_m = 1$ second. The only difference in the narrowband modulation spectrum generated in Section III-B for comparison purposes was $W_a = 30$ ms.

B. Feature Properties and Correlations

Fig. 2 shows the average modulation spectrum per utterance for 4 TIMIT speakers for 3 utterances. These plots show clear peaks in the modulation domain around the fundamental frequencies, which are consistent for each speaker across all utterances by that speaker.

The next step is to analyse the correlations between each modulation spectrum feature with each other modulation spectrum feature and between each modulation spectrum feature and the output speakers. This will help identify the most relevant modulation spectrum features for speaker identification.

1) Correlations With Other Features: Fig. 3 shows the correlation of each modulation spectrum feature with the output speaker using the Spearman's rank correlation coefficient.

This illustrates how strongly correlated the modulation spectrum features are with each other, especially the ones in either the same acoustic frequency band or with the same modulation frequency. It highlights that there are many redundancies in speech and the importance of decorrelating the features for reliable prediction models to be fitted. A similar larger graph for all modulation features is not shown for lack of space.

2) Correlations With Outputs: As this involves finding the correlation between numerical inputs and categorical outputs, the one-way analysis of variance (ANOVA) method is used to calculate the correlation between each input feature and the outputs. The F-statistic F_s for speaker s is given by

$$F_s = \frac{\text{between-class-means covariance}}{\text{intra-class covariance}} = \text{tr}\big(\mathbf{A}_s^{-1}\mathbf{B}\big), \quad (5)$$

where $\mathbf{B} \in \mathbb{R}^{D\times D}$ is the between-class (i.e. between speakers)

TABLE I
CNN STRUCTURE USED

Layer	Filter	Activation	Output Shape
Conv2D	$16, (3, 3)$	ReLU	$(L, 25, 501, 1)$
Conv2D	$16, (3, 3)$	ReLU	$(L, 25, 501, 16)$
MaxPool2D	$(3, 3)$	-	$(L, 8, 167, 16)$
Conv2D	$16, (3, 3)$	ReLU	$(L, 8, 167, 16)$
Conv2D	$16, (3, 3)$	ReLU	$(L, 8, 167, 16)$
MaxPool2D	$(2, 2)$	-	$(L, 4, 83, 16)$
Flatten	-	ReLU	$(L, 5312)$
Dense	-	ReLU	$(L, 1000)$
Dense	-	ReLU	$(L, 1000)$
Dense	-	ReLU	$(L, 512)$
Dense	-	Softmax	$(L, 630)$

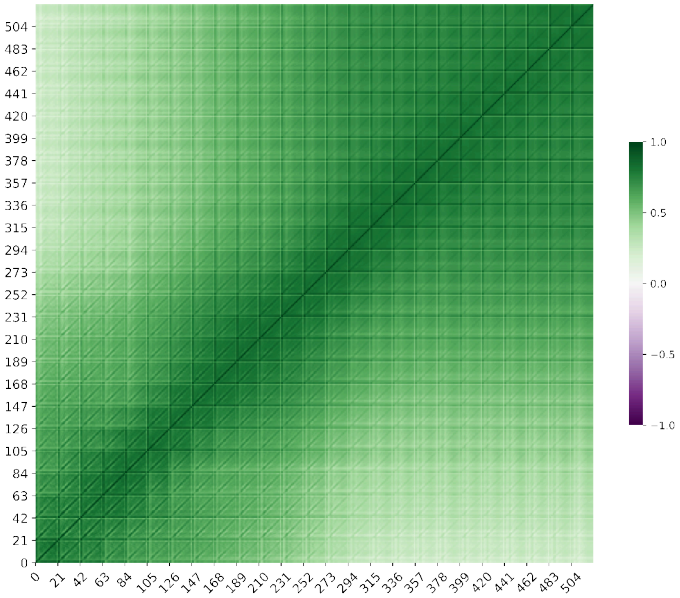

Fig. 3: Correlation heatmap for 0-20 Hz modulation frequencies flattened to $25 \times 21 = 525$ features.

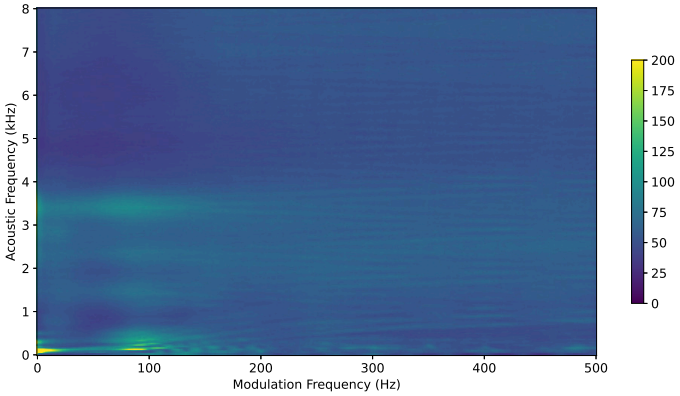

Fig. 5: Narrowband correlations with output speaker labels.

covariance matrix of the means of each class with D as the number of features, $\mathbf{A}_s \in \mathbb{R}^{D\times D}$ is the covariance matrix for speaker $\{s \in \mathbb{Z} : 0 \leq s \leq S-1\}$, D is the total number of features flattened to a 1D array (i.e. the number of dimensions) and tr$(\cdot)$ denotes the trace of a square matrix. This correlation between each input feature (k, h) and the categorical speaker outputs is plotted in Fig. 4(a). The modulation spectrum features from the original TIMIT training set most correlated with the speakers are in acoustic frequency band 1 (indexed from 0) of 166-500 Hz, and the top 20 are all in that band. The most important is in modulation frequency band 107 Hz, with the top 20 all in the range 98 to 120 Hz. Similar correlations occur for the original TIMIT test set, peaking at 115 Hz with all top 20 in acoustic frequency band 1.

The equivalent for the narrowband modulation spectrum is shown in Fig. 5 (peak scaled as top values dominated the plot). The primary peak is around acoustic frequency band 3 (83-117 Hz) and the 0 Hz modulation frequency band; this cluster contains all top 20 values. A secondary peak occurs in the same acoustic frequency band at just below 100 Hz modulation frequencies. Some harmonics in the acoustic domain are also visible. This suggests that peaks occur at speech harmonics in both acoustic and modulation domains.

Fig. 4(b) and (c) plot the Hilbert transform versions. The Hilbert envelope ANOVA scores are higher than those of the amplitude envelope and occur at similar frequencies, albeit peaking at slightly lower modulation frequencies. The temporal fine structure has peaks at higher modulation frequencies and shows the fundamental frequencies more strongly, but the absolute scores are 3 orders of magnitude lower.

TABLE II
SPEAKER IDENTIFICATION PERFORMANCE (ALL IN %)

	Per MF	**Per Utt.**	**Ave. MF**
RF $\mathbf{\Phi}_{AE}$	12.34	27.63	26.20
CNN $\mathbf{\Phi}_{AE}$	29.03	42.40	26.36
CNN $\mathbf{\Phi}_{HE}$	27.97	48.39	32.77
CNN $\mathbf{\Phi}_{IF}$	5.75	12.20	0.69
CNN $\mathbf{\Phi}_{HE}$ and $\mathbf{\Phi}_{IF}$	31.05	49.26	32.17

C. Speaker Identification

Table II shows how well models fitted on $\mathbf{\Phi}$ (either $\mathbf{\Phi}_{AE}$ or $\mathbf{\Phi}_{HE}$) and/or $\mathbf{\Phi}_{IF}$ identify TIMIT test speakers. Results are obtained for (a) each individual modulation frame in the test utterance ("Per MF"), (b) taking the modal prediction for the entire test utterance ("Per Utt.") and (c) making the prediction based on the average $\mathbf{\Phi}(l)$ for each modulation frame in the utterance ("Ave. MF").

Although these performance figures are low, it is not expected that modulation spectrum features alone would perform well on a speaker identification task. Instead, it is expected they be used in combination with other features for highest performance. Furthermore, the small TIMIT dataset is insufficient for training large and complex models on the 630 speakers in it. Nevertheless, it is interesting to study how well modulation spectrum features perform alone.

CNN models perform better than random forest (RF) models, and are quicker to train using less memory/storage. $\mathbf{\Phi}$ is shown to have significant speaker-specific information, and results improve significantly when taking the modal prediction over an utterance rather than looking at each modulation frame prediction in isolation. $\mathbf{\Phi}_{IF}$ contains less speaker-specific information. However, modifying (3) so the first part of the numerator is $\theta_u(m+2, k)$ and interpolating first and last frames gives $\mathbf{\Phi}_{IF} \in \mathbb{R}^{L\times K\times H}$, then changing the denominator to the modulation frame step F_m and using with $\mathbf{\Phi}_{HE}$ as a second CNN channel gives the best results shown in the last row of Table II. This means $\mathbf{\Phi}_{IF}$ relating to the temporal fine structure has additional speaker-specific information.

D. Feature Importances

Fitting the random forest described in Section III-A to the TIMIT data results in the feature importances shown in Fig. 6. Surprisingly, the most important feature was in acoustic frequency band 0 and modulation frequency band 0 Hz, and the 13 most important features were all in the 0 Hz modulation frequency band. After that, clusters occurred at more expected

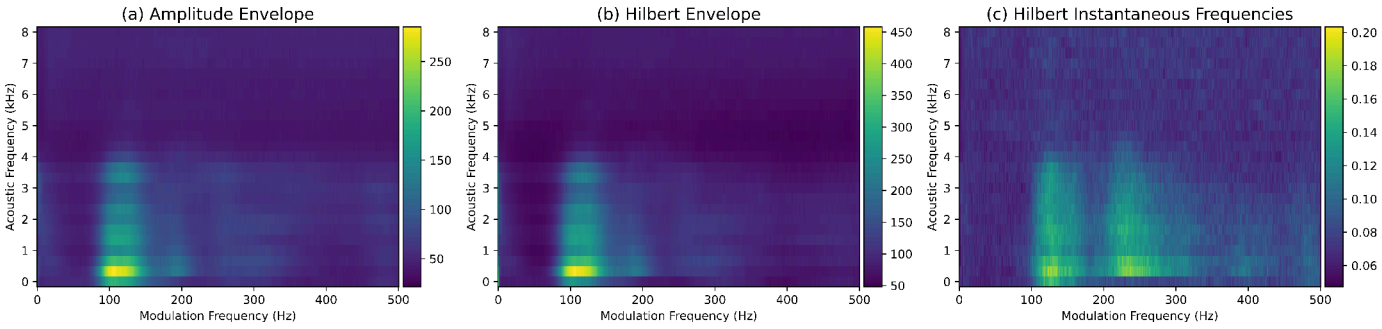

Fig. 4: Correlations of original TIMIT training set modulation spectrum features with output speakers.

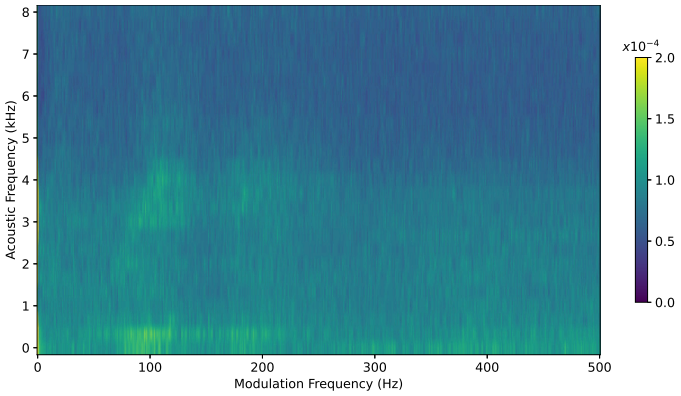

Fig. 6: Feature importances generated by random forest model for TIMIT training set (scale capped at 2×10^{-4}).

places around the speaker fundamental frequencies: features in the range 80-110 Hz modulation frequencies were important, followed in importance by a cluster in the range 180-210 Hz.

E. Modification and Reconstruction

To test how particular parts of the modulation spectrum affect speech, specific elements of $\mathbf{\Phi}_{AE}$ were set to zero and then the signal reconstructed using two stages of inverse STFT (ISTFT). The original speech signal was appended with enough zeros to ensure restoration of the full speech signal. Substantial modifications to the modulation spectrum introduce significant transients at the start of the reconstructed speech signal, which were eliminated by prepending the original speech signal with a modulation frame of zeros. Periodic Hamming windows were used to satisfy the constant overlap-add (COLA) principle [7], [51] for perfect reconstruction.

A range of reconstructed speech signals is shown in [52] for 2 TIMIT speakers with certain acoustic and/or modulation frequencies removed. A formal listening analysis has yet to be conducted on this, though it should be noted that the signals retaining only the 0 Hz modulation frequency band are still surprisingly understandable. Coupled with the graphs showing the importance of the 0 Hz modulation frequency band, it seems that it should not be excluded from speaker identification systems in the way that it is for ASR. As no filter banks are used in this research, the 0 Hz modulation frequency band is the average of each of the modulation frequencies in the range covered, which suggests that specific components from the 0 Hz modulation frequency band should be retained even if filter banks are used in the modulation domain.

IV. Conclusions

The correlations and feature importances presented in this paper have shown that the modulation frequencies of speech most relevant for distinguishing speakers lie in the fundamental frequencies of the speakers rather than the 1-16 Hz range most commonly used in ASR, and that the 0 Hz modulation frequency band contains significant speaker information. For the former, the information in the fundamental frequencies of the speakers is clearly more substantial than F0, so is more than merely an alternative way of expressing F0. For the latter, results from experiments using correlations and random forest models also showed the importance of specific acoustic frequencies in the 0 Hz modulation frequency band. Similarly, reconstructed speech signals without the 0 Hz modulation frequencies were considerably less clear than with them, based on informal listening experiments. Together, this suggests that the 0 Hz modulation frequency band should not be removed altogether, but equally including the entire 0 Hz modulation frequency band is not as helpful as including only acoustic frequency components most relevant for speaker identification.

Results from the experiments with CNN models show that the temporal envelope contains more speaker-specific information than the temporal fine structure; the CNN model trained on temporal envelope features alone gave speaker identification performances of 27.97% for individual modulation frames and 48.39% taking the modal prediction per utterance, whereas the CNN model trained on temporal fine structure features alone gave 5.75% for individual modulation frames and 12.20% for the modal prediction per utterance. However, the CNN trained on both temporal envelope features and temporal fine structure features performed the best; the speaker diarization performance was 31.05% when trained on both, which is a 14.2% improvement on the CNN model trained on temporal envelope features alone, and a similar though less dramatic improvement of the modal prediction per utterance figure to 49.26%, which is a 1.8% improvement. This shows that the temporal fine structure still contains useful additional information that, when used in addition to the temporal envelope, improves speaker discrimination.

The high correlations seen between adjacent acoustic and modulation spectrum features suggests that the use of filter banks to combine frequency bands in both the acoustic and modulation domains should improve performance. However, the results show that in the modulation domain it would be preferable to retain the parts of the 0 Hz modulation frequency band most relevant for speaker identification in addition to such filter banks.

References

[1] H. Beiji, *Fundamentals of Speaker Recognition*. New York: Springer Science & Business Media, 2011.

[2] M.-W. Mak and J.-T. Chien, *Machine Learning for Speaker Recognition*. Cambridge University Press, 2021.

[3] C. S. Greenberg, S. O. Sadjadi, L. Mason, and D. A. Reynolds, "Two decades of speaker recognition evaluation at the National Institute of Standards and Rechnology," Mar. 2020. [Online]. Available: https://www.nist.gov/publications/two-decades-speaker-recognition-evaluation-national-institute-standards-and-technology

[4] O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, and L. Mason, "NIST 2020 CTS speaker recognition challenge evaluation plan," 2020. [Online]. Available: https://www.nist.gov/itl/iad/mig/nist-2020-cts-speaker-recognition-challenge

[5] A. Nagrani, J. S. Chung, J. Huh, A. Brown, E. Coto, W. Xie, M. McLaren, D. A. Reynolds, and A. Zisserman, "VoxSRC 2020: the second VoxCeleb speaker recognition challenge," in *arXiv:2012.06867 [cs, eess]*, Dec. 2020.

[6] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "Third DIHARD challenge evaluation plan," 2020. [Online]. Available: https://dihardchallenge.github.io/dihard3/docs/third_dihard_eval_plan_v1.2.pdf

[7] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*. Upper Saddle River: Pearson Higher Education, Inc., 2011.

[8] N. Morgan, Q. Zhu, A. Stolcke, K. Sönmez, S. Sividas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, Ö. Çetin, H. Bourlard, and M. Athineos, "Pushing the envelope - aside," *IEEE Signal Process. Mag.*, pp. 81–88, 2005.

[9] H. Hermansky, J. R. Cohen, and R. M. Stern, "Perceptual properties of current speech recognition technology," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1968–1985, Sep. 2013.

[10] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2018, pp. 5329–5333.

[11] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2020, pp. 3830–3834.

[12] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "BUT system description to VoxCeleb Speaker Recognition Challenge 2019," in *arXiv:1910.12592 [cs, eess]*, Oct. 2019.

[13] A. O. T. Hogg, C. Evers, and P. A. Naylor, "Speaker change detection using fundamental frequency with application to multi-talker segmentation," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, May 2019, pp. 5826–5830.

[14] R. Baken and R. Orlikoff, *Clinical Measurement of Speech and Voice*, 2nd ed. Cengage Learning, Dec. 1999.

[15] Y. Cheng and H. Leung, "Speaker verification using fundamental frequency," in *Proc. Int. Conf. on Spoken Language Process. (ICSLP)*, 1998, pp. 161–164.

[16] H. Hermansky, "History of modulation spectrum in ASR," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2010, pp. 5458–5461.

[17] F. Apoux, S. E. Yoho, C. L. Youngdahl, and E. W. Healy, "Role and relative contribution of temporal envelope and fine structure cues in sentence recognition by normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 134, no. 3, pp. 2205–2212, Sep. 2013.

[18] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1053–1064, Feb. 1994.

[19] ——, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Am.*, vol. 95, no. 5, pp. 2670–2680, May 1994.

[20] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 1997, pp. 1647–1650.

[21] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition," in *Proc. Eur. Conf. on Speech Communication and Technol. (EUROSPEECH)*, 1997, pp. 1079–1082.

[22] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Syllable intelligibility for temporally-filtered LPC cepstral trajectories," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2783–91, May 1999.

[23] L. Varnet, M. C. Ortiz-Barajas, R. G. Erra, J. Gervain, and C. Lorenzi, "A cross-linguistic study of speech modulation spectra," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 1976–1989, Oct. 2017.

[24] A. L. Giraud, C. Lorenzi, J. Ashburner, J. Wable, I. Johnsrude, R. Frackowiak, and A. Kleinschmidt, "Representation of the temporal envelope of sounds in the human brain," *Journal of Neurophysiology*, vol. 84, no. 3, pp. 1588–1598, Sep. 2000.

[25] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP J. on Advances in Signal Process.*, vol. 2003/310290, Dec. 2003.

[26] S. M. Schimmel, L. E. Atlas, and K. Nie, "Feasibility of single channel speaker separation based on modulation frequency analysis," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2007, pp. IV–605–IV–608.

[27] F.-G. Zeng, K. Nie, G. S. Stickney, Y.-Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proc. National Academy of Sciences*, vol. 102, no. 7, pp. 2293–2298, Feb. 2005.

[28] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 16, no. 6, pp. 1097–1111, Aug. 2008.

[29] D. Gowda, R. Saeidi, and P. Alku, "AM-FM based filter bank analysis for estimation of spectro-temporal envelopes and its application for speaker recognition in noisy reverberant environments," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2015, pp. 1166–1170.

[30] X. Teng, G. B. Cogan, and D. Poeppel, "Speech fine structure contains critical temporal cues to support speech segmentation," *NeuroImage*, vol. 202, no. 116152, Nov. 2019.

[31] B. C. J. Moore, "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people," *Journal of the Association for Research in Otolaryngology*, vol. 9, no. 4, pp. 399–406, Dec. 2008.

[32] ——, "Interference effects and phase sensitivity in hearing," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 360, no. 1794, pp. 833–858, May 2002.

[33] K. Vijayan, P. Raghavendra Reddy, and K. Sri Rama Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication*, vol. 81, pp. 54–71, Jul. 2016.

[34] S. Sheft, M. Ardoint, and C. Lorenzi, "Speech identification based on temporal fine structure cues," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 562–575, Jul. 2008.

[35] I. J. Moon and S. H. Hong, "What is temporal fine structure and why is it important?" *Korean Journal of Audiology*, vol. 18, no. 1, pp. 1–7, Apr. 2014.

[36] S. van Vuuren and H. Hermansky, "On the importance of components of the modulation spectrum for speaker verification," in *Proc. Int. Conf. on Spoken Language Process. (ICSLP)*, 1998.

[37] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[38] T. Kinnunen, "Joint acoustic-modulation frequency for speaker recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, vol. 1, 2006, pp. I–665–I–668.

[39] T. Kinnunen, K. A. Lee, and H. Li, "Dimension reduction of the modulation spectrogram," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2008.

[40] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Contributions of temporal cue on the perception of speaker individuality and vocal emotion for noise-vocoded speech," *Acoustical Science and Technol.*, vol. 39, no. 3, pp. 234–242, May 2018.

[41] O. Vinyals and G. Friedland, "Modulation spectrogram features for improved speaker diarization," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2008, pp. 630–633.

[42] D. Vijayasenan, "An information theoretic approach to speaker diarization of meeting recordings," Ph.D. dissertation, École polytechnique fédérale de Lausanne, Lausanne, Switzerland, 2010.

[43] D. Vijayasenan and F. Valente, "DiarTk: an open source toolkit for research in multistream speaker diarization and its application to meetings recordings," in *Proc. Conf. of Int. Speech Commun. Assoc. (INTERSPEECH)*, 2012, pp. 2170–2173.

[44] D. Vijayasenan, F. Valente, and H. Bourlard, "Multistream speaker diarization of meetings recordings beyond MFCC and TDOA features," *Speech Communication*, vol. 54, no. 1, pp. 55–67, Jan. 2012.

[45] D. Sharma, A. O. T. Hogg, Y. Wang, A. Nour-Eldin, and P. A. Naylor, "Non-intrusive POLQA estimation of speech quality using recurrent neural networks," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2019.

[46] Y. Wang, "Speech enhancement in the modulation domain," Ph.D. dissertation, Imperial College London, London, 2015.

[47] T. H. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 90–100, 2010.

[48] M. Diez, L. Burget, F. Landini, S. Wang, and H. Černocký, "Optimizing Bayesian HMM based x-vector clustering for the second DIHARD speech diarization challenge," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2020, pp. 6519–6523.

[49] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA-TIMIT: acoustic-phonetic continuous speech corpus CD-ROM," Nat. Inst. of Standards and Tech. (NIST), Interagency/Internal Report 4930, 1993.

[50] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, Massachusetts, USA: The MIT Press, 2012.

[51] C. Borß and R. Martin, "On the construction of window functions with constant-overlap-add constraint for arbitrary window shifts," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2012, pp. 337–340.

[52] S. W. McKnight, "Reconstructed speech signals removing specific modulation spectrum components," 2021. [Online]. Available: https://swm1718.github.io/ModulationSpectrumAudio/