NTIA Technical Memorandum 24-574

A Powerful, Fixed-Size Modulation Spectrum Representation for Perceptually Consistent Speech Evaluation

Stephen D. Voran
Jaden Pieper

Technical Memorandum

NTIA Technical Memorandum 24-574

A Powerful, Fixed-Size Modulation Spectrum Representation for Perceptually Consistent Speech Evaluation

Stephen D. Voran
Jaden Pieper

U.S. DEPARTMENT OF COMMERCE

Alan Davidson
Assistant Secretary of Commerce for Communications and Information
National Telecommunications and Information Administration

September 2024

DISCLAIMER

Certain commercial equipment and materials are identified in this report to specify adequately the technical aspects of the reported results. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the material or equipment identified is the best available for this purpose.

CONTENTS

FIGURES

TABLES

A POWERFUL, FIXED-SIZE MODULATION SPECTRUM REPRESENTATION FOR PERCEPTUALLY CONSISTENT SPEECH EVALUATION

Stephen D. Voran, Jaden Pieper[1]

We develop the wideband fixed-size modulation spectra (FMS) and show that they contain the necessary information to perform perceptually consistent evaluation of speech. We compare FMS with the already established frame-based modulation spectra as representations for estimating speech quality and speech naturalness. We feed the two representations into equally sized, relatively small, fully connected networks for five proof-of-concept experiments and find that the two representations perform similarly when estimating speech quality and speech naturalness. But the FMS representation captures an entire speech file into a fixed size representation, which means that additional temporal processing is neither needed nor possible. This is in contrast to the frame-based representations, where additional processing can either destroy information (e.g., averaging over time) or lead to more complex and difficult to train networks.

We also demonstrate that when speech quality changes within a speech file, FMS has another distinct advantage which is to be able to efficiently and reliably identify different situations in a way that is not well-addressed by the frame-based approach. Our experiments include more than 274 hours of speech in two languages. This speech is contained in 139,000 speech files and there is a subjective score for each file. File lengths range from 0.6 to 28 seconds and five different sample rates are present. Supporting software is available at `https://github.com/NTIA/fms`.

Keywords: Auditory perception, fixed-size modulation spectra (FMS), machine learning, mel spectrum, modulation spectrum, speech naturalness, speech quality, time-varying speech quality

[1]The authors are with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, Colorado 80305.

1 INTRODUCTION

As mobile and fixed telecommunications options continue to proliferate and evolve, users commonly experience dramatic variations in speech quality due to different devices, connections, and acoustic environments. Accurate real-time measurements of received speech quality without reference to transmitted speech are invaluable. Thus no-reference (NR) speech quality estimation, often enabled by machine learning (ML), is an active research area. ML can extract powerful frame-based speech-quality features directly from waveforms (e.g., [1]) but this can be computationally intensive. Hand-crafted frequency-domain features can be an efficient alternative. These often involve the short-time Fourier transform (STFT) and thus are also frame-based, meaning an individual set of features is produced for each 10 to 30 ms frame of speech (e.g., [2]–[4]).

No matter the source, a frame-based feature sequence requires some sort of processing that extracts a single speech quality value from a sequence of frame-level results. Averaging is an easy solution [2], [5], [6], but is not fully consistent with human perception [7], [8]. ML can leverage attention mechanisms and long short-term memory (LSTM) structures to potentially learn more intricate solutions, but these can dramatically increase architecture complexity. This memorandum demonstrates that the modulation spectrum (MS) can extract the needed information, and, when properly implemented, can also bypass the need for processing of frame-level results.

An MS describes how the power at different acoustic frequencies changes over time. The frame-based narrowband MS has been successfully used to estimate narrowband speech quality in a variety of contexts (e.g., [9]–[12]) but as a frame-based feature, it requires a processor that reduces the sequence of per-frame results to a single value, and this has typically meant some type of averaging MS over time. In addition, current approaches often include assumptions about desirable and undesirable modulation frequencies, thresholding, and additional averaging that are not always fully aligned with human perception. In some cases (e.g. [13]–[16]) ML has been invoked to perform data reduction that reduces, but does not eliminate, the need for potentially harmful averages and assumptions.

In this memorandum we present wideband fixed-size modulation spectra (FMS, singular or plural) that generate global representations of entire signals (as opposed to a frame-based approach). Thus it produces a fixed-size representation for a wide range of signal lengths. It uses the perceptually motivated mel frequency scale [17] to cover the wideband or fullband audio frequency range (nominally extending to 8 or 20 kHz, respectively) rather than the narrowband range which nominally extends to just 4 kHz. It presently accommodates six sample rates and is easily extended to others. We allow the entire MS to be used by subsequent ML — there are no averaging, thresholding, or processing assumptions. Our code is available at `https://github.com/NTIA/fms`.

In the following we describe the problem and previous MS-based work in more detail. We then develop FMS, clearly contrast it with previous approaches, and provide example results. In Section 4 we offer proof-of-concept experiments that demonstrate the utility of FMS in emulating human perceptual tasks. These experiments add lightweight networks to FMS in order to emulate human perception and judgement for rating speech quality and speech naturalness. The FMS results are similar to results from frame-based modulation spectra when we employ identical lightweight

networks for the estimation task.

In Section 5 we show that FMS has a unique ability to detect different situations when speech quality varies in time. This proof-of-concept work suggests that FMS has a strong innate advantage for estimating time-varying speech quality. A full demonstration of this is not possible until sufficient subjective scores are available for time-varying speech quality recordings.

While prior MS speech-quality work used no more than a few thousand speech files at most, generous sharing of crowdsourced ratings allows us to present results from over 139,000 speech files and subjective scores (before augmentation). This is more than 274 hours of speech and both English and Chinese are represented. The results allow us to conclude that FMS are compact, fixed-size representations for entire speech files that need only very simple networks to produce human-like judgments of speech.

2 FRAME-BASED SPEECH QUALITY FEATURES

The first step for any machine learning method for speech quality estimation involves feature extraction. This can be done through digital signal processing (DSP) on raw speech inputs, or through the first learned component in the network, or in combinations of these two (e.g., [2], [4], [18]–[20]). In principle, raw speech is the ideal input, as no information is lost and the network has the ability to learn the most meaningful representation of the speech. However, this approach requires a vast amount of training data and can require complex, large network components in order to learn interesting representations.

DSP can be applied to the speech prior to entering the network in order to generate a feature rich representation, often in the frequency domain using the STFT. Further, by aggregating through meaningful filter banks (e.g., computing mel spectra), we can create lower-dimensional, perceptually relevant representations of the speech.

An alternative approach can utilize the frame-based modulation spectrum (MS). In [21] it was observed that temporal envelopes are related to speech quality and intelligibility and those observations are supported by earlier work in [22]–[24]. Also, [21] suggests the MS as a suitable basis for detailed analysis of temporal envelopes. The MS had been developed earlier (and named modulation *spectrogram*) as a perceptually grounded signal representation for speech enhancement [25]. The no-reference telecommunications speech-quality estimator called ANIQUE [9] uses the MS to separate modulations below 30 Hz that are often associated with speech production from other modulations that may be caused by impairments to the speech signal. This separation feeds articulation-to-nonarticulation ratios (ANRs) which are accumulated across 23 critical bands and then combined across frames to produce a single quality estimate for an entire speech file. In ANIQUE+ a small neural network is used to combine per-critical band articulation power, nonarticulation power, and ANR ($3 \times 23 = 69$ values) to a single value for each frame [13]. In [15], the mean, variance, skewness, and kurtosis of these 69 values are computed over frames and the resulting 276 statistics are combined with additional parameters and used to predict speech quality. The MS also provides the basis for the speech-to-reverberation modulation energy ratio (SRMR) used to estimate speech quality and speech intelligibility in [10]–[12]. ML was used to further advance this work in [14], [16].

Extracted features typically represent time through frame-based representations, where a set of features are extracted from each time-frame or window across a speech signal. Frame-based feature extraction is powerful and easily accommodates speech signals of different lengths. But it must be followed by appropriate processing to extract a value from a sequence. Humans do not average short-term impressions of speech quality to arrive at longer term impressions. Instead, larger and more frequent quality variations reduce long-term quality [7]. In addition, a long-term rating is typically lower-bounded by the minimum of short-term ratings and upper-bounded by the average of short-term ratings, and the recency effect can also be a factor [7], [8]. But speech quality estimators often simply average per-frame quality estimates to arrive at a single global quality estimate [2], [5], [6].

Networks can use recurrent layers (e.g., LSTM) and attention mechanisms to learn frame-based

relations and aggregations that are more nuanced and perceptually consistent than a simple average [18]. However, these layers can dramatically increase the complexity and size of the network. In addition, since time-varying impairments are often under-represented compared to constant impairments, it may be difficult to enable these layers to properly learn meaningful and extendable relationships. Baking meaningful time relationships into the speech features up front can potentially mitigate these problems and allow for smaller, simpler networks with perceptually consistent temporal behavior. This motivates a different, light-weight approach for feature extraction that does not rely on frame-based representations and the associated aggregations that can be at odds with human speech perception.

3 FIXED-SIZE MODULATION SPECTRUM

To our knowledge, all prior uses of MS for speech quality estimation have used a narrowband, frame-based MS. In addition to the inherently required aggregation of samples into perceptual acoustic bands and perceptual modulation bands, these have also imposed additional forms of grouping, averaging, or thresholding out of necessity. Here we develop a frame-less, wideband, fixed-size MS that does not impose any additional processing or assumptions beyond the acoustic and modulation filter banks.

In the following we require the input speech signal to have a duration of at least three seconds. This is not a significant restriction, since the speech signals used in telecommunications testing almost always meet this requirement; and when they fall a bit short, zero padding can be used to meet the requirement. This duration requirement stems from our desire to capture modulation frequencies as low as 0.25 Hz because these modulations can capture attributes of time-varying speech quality.

Given a sequence of audio samples x_i ($i = 0$ to $N_a - 1$), we first calculate frame-based mel spectra using the normalized periodic Hamming window with 16 or 17.4 ms duration as shown in Table 1. We then zero pad each frame to twice its original length and apply the STFT. The zero padding is needed to achieve the desired spectral resolution. These STFT results are then aggregated using a mel-scale filter bank [17]. Full details of this filter bank are provided in Appendix A.

This yields an $N_{\text{win}} \times N_{\text{mel}}$ representation of the signal, where N_{win} is determined by the sample rate of the audio signal, f_s, and the stride, N_s, used in the STFT. N_{mel} describes the number of mel-frequency bands considered. Let $P_{i,w}$ describe the square root of the power in mel band i, for window w such that $P_{i,\cdot}$ is a time history of the square root of the power in mel band i and that history (or envelope) is sampled at f_s/N_s Hz.

For any f_s we select N_s to keep this new sample period approximately constant ($N_s/f_s \approx 2$ ms) to enable consistent modulation spectrum analyses across f_s values (see Table 1). We perform spectral analysis on each envelope using the symmetric unnormalized Hamming window and the DFT:

$$\Gamma_{i,k} = \sum_{w=0}^{N_{\text{win}}-1} P_{i,w} \left(0.54 - 0.46 \cos\left(\frac{2\pi w}{N_{\text{win}} - 1} \right) \right) e^{-j2\pi kw/N_{\text{win}}},$$
$$i = 0 \text{ to } N_{\text{mel}} - 1, \;\; k = 0 \text{ to } N_{\text{win}} - 1. \tag{1}$$

$\Gamma_{i,k}$ are complex envelope modulation spectrum values for mel band i. The index k corresponds to modulation frequency via $f_{\text{mod}} = (f_s/N_s) \times k/(N_{\text{win}} - 1)$.

Following the successful precedent set in [9] and [10], we aggregate these values in a logarithmic fashion. But in a departure from precedent, FMS uses uniform rather than triangular weighting (or modulation filters). This choice is motivated by the sparsity of spectral samples at the low frequency end of our analysis range (i.e., 0.25 Hz). This sparsity precludes smooth triangular weighting but does not interfere with uniform weighting.

Table 1. FMS parameters selected to produce 32 identical mel bands between DC and 8 kHz for six sample rates, f_s. N_s is chosen so that $N_s/f_s \approx 2$ ms for each sample rate. The constraint that mel bands be equally sized can cause the upper frequency limit u_f to be less than Nyquist frequency.

f_s (kHz)	Win. Len. (samples)	Win. Len. (ms)	N_s (samples)	Stride Duration (ms)	u_f (kHz)	N_{mel}
16	256	16.000	32	2.000	8.000	32
24	384	16.000	48	2.000	11.108	36
32	512	16.000	64	2.000	15.326	40
48	768	16.000	96	2.000	22.778	45
22.05	384	17.415	44	1.995	10.240	35
44.1	768	17.415	88	1.995	21.052	44

This uniformly weighted, logarithmically spaced aggregation of modulation spectrum values can be viewed as a filter bank operation and is fully described in Appendix B. That is, we use the bank of N_{mod} rectangular filters $\Phi_{k,m}$ defined in Appendix B to produce modulation spectrum results.

The filters $\Phi_{k,m}$ aggregate spectral magnitudes and spectral phases:

$$\Psi^M_{i,m} = \sum_{k=0}^{\lfloor N_{\text{win}}/2 \rfloor} |\Gamma_{i,k}| \Phi_{k,m}, \quad \Psi^P_{i,m} = \sum_{k=0}^{\lfloor N_{\text{win}}/2 \rfloor} \angle(\Gamma_{i,k}) \Phi_{k,m},$$
$$i = 0 \text{ to } N_{\text{mel}} - 1, \quad m = 0 \text{ to } N_{\text{mod}} - 1, \tag{2}$$

where $\lfloor \cdot \rfloor$ is the floor function and $\angle(\cdot)$ extracts the angle of the complex argument. Together, $\Psi^M_{i,m}$ and $\Psi^P_{i,m}$ form the fixed-size modulation spectrum. We use $N_{\text{mod}} = 11$ and the associated 11 modulation frequencies are nominally 0, 0.25, 0.50, 1.0, 2.0, ..., and 128 Hz.

The length of the speech signal and the stride, N_s, determine the number of windows produced, N_{win}. This in turn determines the envelope modulation spectral resolution in $\Gamma_{i,k}$. This resolution is removed in the aggregations given in (2), resulting in modulation spectra $\Psi^M_{i,m}$ and $\Psi^P_{i,m}$ that have fixed-size, (size $N_{\text{mel}} \times N_{\text{mod}}$), *independent* of the length of the speech signal.

Table 1 shows FMS parameters we selected for uniformity across six sample rates. For each sample rate we calculated N_{mel} and an upper frequency limit u_f such that the steps in Appendix A yield 32 identical uniform mel bands between DC and 8 kHz. The window stride N_s equates to about 2 ms and this becomes the envelope sample period (equivalent to an envelope sample rate of 500 Hz). We selected this value so that our analyses include modulation frequencies up to 250 Hz.

FMS and the ways that we use them are very different from frame-based narrowband MS and the ways they have been used:

- FMS are global (not frame-based) in the sense that information from an entire speech file is represented in a fixed-size MS. This allows FMS to provide per-signal results without restricting signals to a fixed length (e.g., [19], [26]) or explicit temporal processing (e.g.,

[14], [18], [20], [27]). This is significant as the work presented in Section 4 uses signals with durations from 3.0 (after zero padding) to 28.0 seconds and FMS represent them all without losing relevant information. This approach fits best with file-based (not stream-based) workflows.

- FMS support wideband (and fullband) audio analysis.
- FMS use all information, and all aggregation is perceptually grounded. They do not average over frames, apply thresholds, and do not invoke any assumptions about the meaning or importance of different modulation bands.

4 EXAMPLE FMS APPLICATIONS

We now demonstrate that FMS contain sufficient information to emulate human ratings of speech quality and speech naturalness. We do this first by exploring FMS representations of example speech files, and then with five proof-of-concept experiments that use extremely simple, relatively small networks.

4.1 Example FMS structure

FMS create a rich representation of speech, and speech signal impairments are reflected in an FMS through intricate interactions. Figure 1 shows results for an example speech segment (4.1 second duration, female talker saying "The wreck occurred by the bank on Main Street. The doorknob was made of bright clean brass.") The figure also shows results for three impaired versions of that original speech. These include impairment by non-stationary coffee shop noise at 0 dB SNR, impairment by significant reverberation, and impairment by the muting of 20 ms segments at random locations. The original speech MS provides a reference. It shows that modulation power drops off steadily as modulation frequency increases and that higher acoustic frequencies carry less modulation. While this trend is still seen when noise, reverb, and muting impair the speech, it is clear that these impairments also cause noticeable differences in the MS.

Figure 2 showcases the differences and general trends in these differences. That figure displays mathematical differences (impaired log MS minus reference log MS) for each mel band and modulation band. In this example, noise has reduced MS power at the middle modulation frequencies for most mel bands, but the lowest mel band shows increased power at all modulation frequencies. The example reverberation has caused a marked decrease in MS power for just two of the middle modulation bands, and an increase in the lower mel bands at the higher modulation frequencies. Finally, the random muting example produced dramatic decreases in modulation band 2, but only at higher mel bands, and dramatic increases in modulation band 8, but only at lower mel bands. These examples make it clear that different impairments cause unique deviations in the MS, and thus suggests that the MS could be analyzed to understand types and severities of impairments.

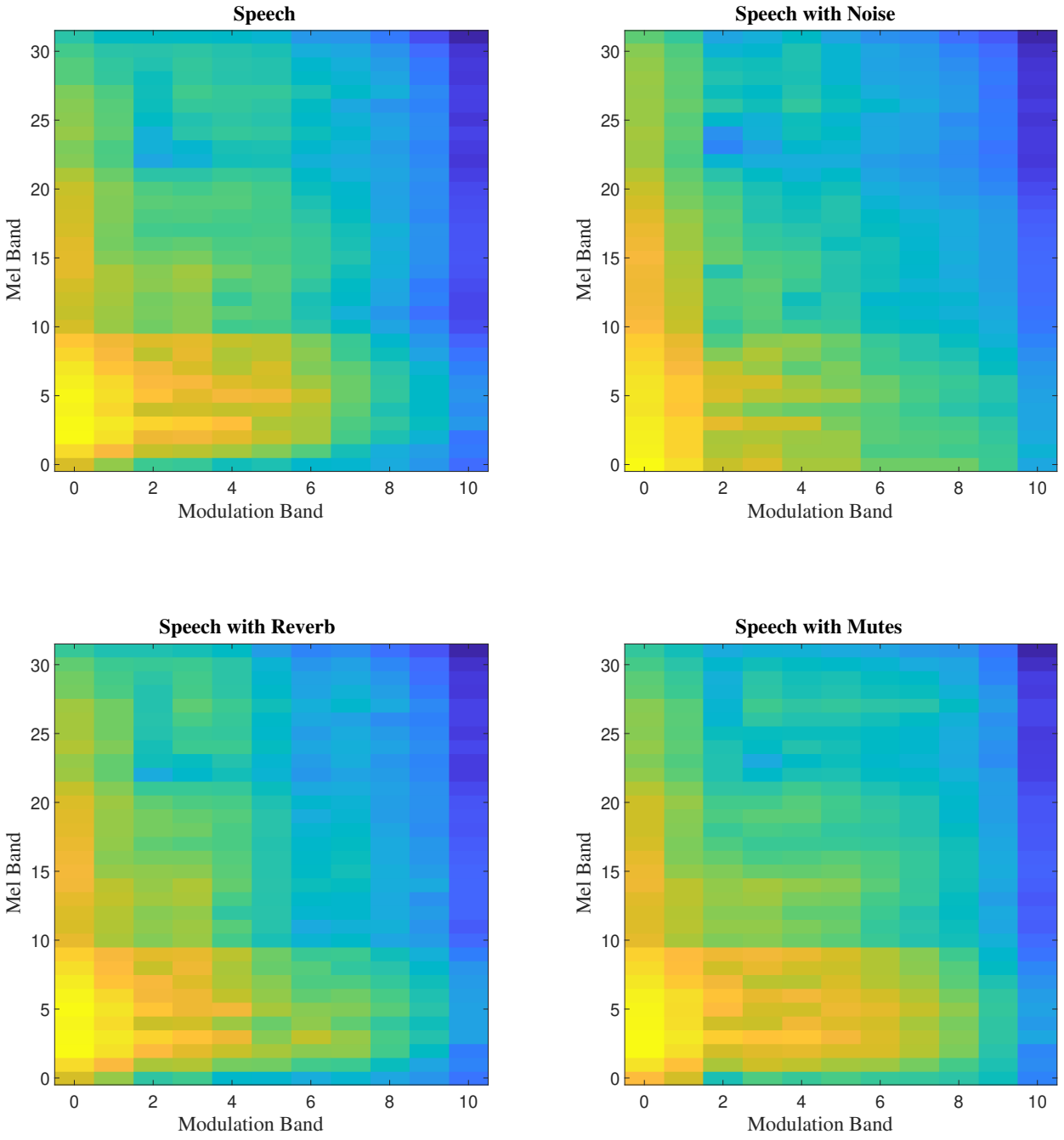

Figure 1. Images showing example log-scale modulation spectra for speech, speech with noise, speech with reverb, and speech with muting (dark blue for smallest values and bright yellow for largest values).

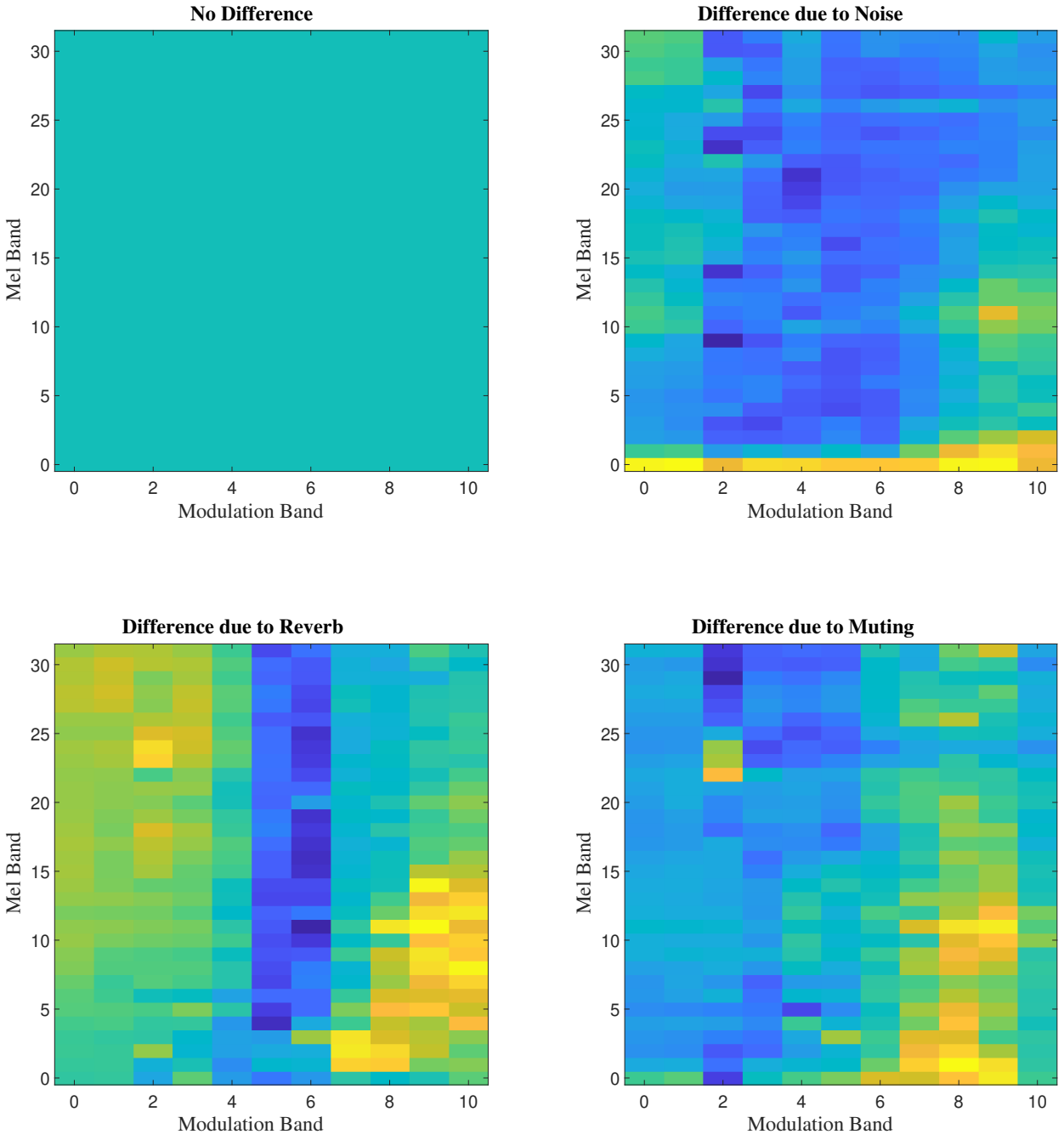

Figure 2. Images showing differences in log-scale modulation spectra arising when speech is impaired by noise, reverb, or muting (dark blue for largest power decreases and bright yellow for largest power increases).

Table 2. Dataset summary. Values shown for Tencent, VCC2018, and NISQA include a doubling produced by data augmentation.

Dataset	Language	f_s (kHz)	Lengths (s)	Rating Task	Training (files)	Validation (files)	Testing (files)	Total Hours
IUB [20]	English	16	2.0 – 7.8	Quality	30,600	1800	3600	38.5
Tencent [28]	Chinese	16, 32, 44.1, 48	5.0 – 28.0	Quality	19,958	1156	2312	47.0
VCC2018 [29]	English	16, 22.05	0.6 – 7.9	Naturalness	34,986	2058	4116	37.6
NISQA [30]	English	48	4.5 – 12.0	Quality	21,250	1250	2500	61.3
PSTN [31]	English	8	10.0	Quality	49,904	2935	5870	163.0

4.2 FMS and frame-based MS as features for speech quality estimation

We ran experiments to compare FMS with conventional frame-based MS. The experiment task is to estimate subjective speech quality scores for four databases and subjective speech naturalness scores for a fifth database. FMS and frame-based MS are used as inputs to identical extremely simple, relatively small networks and those networks are trained to produce the desired targets. We then report performance using set-aside testing data from each database. We expect that the results here form a lower bound for what could be achieved with more insightful and complex networks. The intent of these experiments is not to show optimal solutions, but rather to demonstrate that FMS carry relevant speech quality and speech naturalness information similar to that carried in frame-based MS. Key attributes of the five datasets are summarized in Table 2.

Across the first four datasets, the lowest sample rate is 16 kHz, so the smallest number of mel bands available is 32. For consistency, experiments on these datasets use only these 32 lowest mel bands ($N_{\text{mel}} = 32$) covering 0 to 8 kHz. To continue with the consistent size approach, we seek to use 32 mel bands on the PSTN dataset as well. Since the PSTN dataset is narrowband (sample rate is $f_s = 8$ kHz and Nyquist frequency is 4 kHz), those 32 mel bands are located between DC and 4 kHz (i.e., $u_f = 4000$ Hz in Appendix A). Other than this necessary departure, the FMS calculations for the PSTN database match those of the other databases by using a 16 ms window and a 2 ms stride.

Each FMS has $N_{\text{mel}} \times N_{\text{mod}} = 32 \times 11 = 352$ magnitude values and 352 phase values. The magnitude values are used to construct a length 352 column vector $\boldsymbol{x_0}$, where the individual entries of $\boldsymbol{x_0}$ are given by

$$x_{32m+i} = \log_{10}(\Psi^M_{i,m}), \;\; i = 0 \text{ to } 31, \;\; m = 0 \text{ to } 10. \tag{3}$$

In general, $\Psi^P_{i,m}$ can provide important additional information. For example, $\Psi^M_{i,m}$ is nearly identical if we place a short noise burst at the start of a speech file or at the end of the speech file, but $\Psi^P_{i,m}$ changes dramatically based on the location of the noise burst. In this example $\Psi^P_{i,m}$ carries impairment location information that could be needed in some applications. We provide a much more formal and quantitative demonstration of the utility of $\Psi^P_{i,m}$ in Section 5. But for the speech data and rating tasks used in this current section, we have found that $\Psi^P_{i,m}$ does not add value, so we form the network input $\boldsymbol{x_0}$ from the logarithm of $\Psi^M_{i,m}$ only.

In all experiments the network first removes the global mean of the training set, and then applies four fully connected layers with ReLU activation:

$$\boldsymbol{x}_{i+1} = \max(W_i \boldsymbol{x_i} + \boldsymbol{b_i}, \boldsymbol{0}), \quad i = 0 \text{ to } 3. \tag{4}$$

In the i^{th} layer, the weight matrix W_i has size $N_i \times N_{i+1}$ and the bias column vector $\boldsymbol{b_i}$ has length N_{i+1}. Layer zero maps the length 352 input vector of FMS values to a length 256 internal representation, so $N_0 = 352$ and $N_1 = 256$. Layers one and two continue to process length 256 internal representations, so $N_2 = N_3 = 256$. And layer three (the fourth and final layer) maps the resulting internal representation to a final scalar value ($N_4 = 1$).

Our experiments are made possible through generous sharing of crowdsourced subjective test results. Crowdsourced tests lack many of the controls used in laboratory tests, but very large sample sizes, simple cross-checks, and data cleaning can more than compensate for that lack of control [32]. Key attributes of the datasets are shown in Table 2 and further details are provided in 4.3 below.

For each experiment, we used 85% of the data and the Adam optimization algorithm to train weights and biases. We used 5% of the data for validation and terminated training when performance on this data was consistently worse than on training data. We used the previously unseen 10% of the data for testing. We ran 10 repetitions of each experiment, and present averaged results for unseen test data on the FMS row in Table 3 and Table 4. All targets are mean opinion scores (MOS) on a scale between 1 and 5.

The FMS results can be compared with frame-based MS results, also shown in Table 3 and Table 4. We constructed the frame-based MS to match FMS, where possible. Thus the frame-based approach uses the same mel spectra as FMS and $N_{\text{mel}} = 32$. But instead of analyzing the entire time-history of the mel spectra as with FMS, we analyze each individual 256 ms (128 sample) window as in [9], [10]. We advance this window with a stride of 32 ms (16 samples) as in [10]. This is made precise by adding the frame index f to (1) and setting $N_{\text{win}} = 128$:

$$\Gamma_{i,k,f} = \sum_{w=0}^{N_{\text{win}}-1} P_{i,16f+w}\big(0.54 - 0.46\cos\left(\frac{2\pi w}{N_{\text{win}}-1}\right)\big)e^{-j2\pi kw/N_{\text{win}}},$$
$$i = 0 \text{ to } N_{\text{mel}}-1, \;\; k = 0 \text{ to } N_{\text{win}}-1, \;\; f = 0 \text{ to } N_{\text{frm}}-1. \tag{5}$$

Consistent with [9] and [10], we then aggregate these values using logarithmically spaced triangular filters $\hat{\Phi}_{k,m}$ fully defined in Appendix C. One filter is at DC, and the others cover the range from 4 to 128 Hz which is again similar to the work in [9]. In order to match FMS, we use $N_{\text{mod}} = 11$ filters total, so one filter is at DC and 10 filters cover the range from 4 to 128 Hz.

$$\Psi^M_{i,m,f} = \sum_{k=0}^{64} |\Gamma_{i,k,f}|\hat{\Phi}_{k,m}\,, \quad \Psi^P_{i,m,f} = \sum_{k=0}^{64} \angle(\Gamma_{i,k,f})\hat{\Phi}_{k,m}\,,$$
$$i = 0 \text{ to } N_{\text{mel}}-1, \quad m = 0 \text{ to } N_{\text{mod}}-1, \;\; f = 0 \text{ to } N_{\text{frm}}-1, \tag{6}$$

where $\angle(\cdot)$ extracts the angle of the complex argument.

We then average $\Psi^M_{i,m,f}$ over all frames as in [10] to produce a single MS value for each of the 32 acoustic frequencies and each of the 11 modulation frequencies. Consistent with FMS, the logarithm of the resulting 352 values become the elements of the feature vector $\boldsymbol{x_0}$:

$$x_{32m+i} = \log_{10}\left(\frac{1}{N_{\text{frm}}}\sum_{f=0}^{N_{\text{frm}}-1}\Psi^M_{i,m,f}\right), \;\; i = 0 \text{ to } 31, \;\; m = 0 \text{ to } 10. \tag{7}$$

The network architecture and training details for the frame-based approach match those for FMS exactly.

Table 3 shows correlations between the estimators and the subjective scores. The table makes it clear that the estimation problem is harder in some datasets (e.g., VCC2018) and easier in others (e.g., IUB). This is caused by several factors, including the types, ranges, and numbers of different impairments found in each database, along with the methodology and number of subjects used to obtain a mean subjective score for each file. Across this wide range of estimation problems we see that the FMS correlation is between 0 and 5% lower than that of frame-based MS. On average, the correlation is reduced by 2% when FMS is used in place of the frame-based approach.

Table 4 is analogous to Table 3 but it provides RMS estimation errors in place of correlations. Here we see that the FMS approach increases the RMS error slightly over that of the frame-based approach. That increase ranges from 1 to 7% with an average value of 4%.

We earlier suggested that performing speech quality estimation by using FMS and frame-based MS as inputs to extremely simple, relatively small networks might illustrate a lower bound for what could be achieved with more insightful and complex networks. Here we offer one comparison with the somewhat earlier work reported in [2]. In that study, speech quality was estimated by passing a Hertz-scale magnitude spectrum through a range of different ML architectures built from various combinations of CNN and BLSTM blocks. Many of these architectures were significantly more complex than the networks we have used here. While the training, validation, and test split ratios differ, the results in [2] for the VCC2018 database include correlations ranging from 0.25 to 0.64. The values given in Table 3 for VCC2018 (0.61 and 0.62) are near the top end of this correlation range. The MSE values given in [2] equate to RMSE values ranging from 0.70 to 0.90. The values given in Table 4 for VCC2018 (0.69 and 0.70) are at the very bottom of this range. The values in [2] are certainly affected by the smaller amount of training data used for the networks, but these results still suggest that MS contain rich speech information. In particular the MS captures temporal relationships directly, obviating the need for sophisticated network elements as typically required with networks using Hertz-scale magnitude spectra as inputs.

Table 3. Linear (Pearson) correlation coefficients for unseen test data.

	IUB	Tencent	VCC2018	NISQA	PSTN
FMS	0.97	0.85	0.61	0.74	0.74
frame-based	0.97	0.86	0.62	0.78	0.75

Table 4. RMS errors for unseen test data.

	IUB	Tencent	VCC2018	NISQA	PSTN
FMS	0.21	0.62	0.70	0.74	0.59
frame-based	0.20	0.60	0.69	0.69	0.57

4.3 Dataset Details

4.3.1 Speech quality for IUB dataset

The Indiana University Bloomington (IUB) dataset [20] contains high-quality speech from close-talking microphones as well as lower-quality speech from more distant microphones. Distant microphones capture more background noise (SNRs range from -10 to +11 dB) and reverberation (speech-to-reverberation ratios range from -5 to +4 dB). A portion of the files were low-pass filtered at 3.4 kHz to create anchor conditions for subjective testing. The crowdsourced speech-quality ratings were filtered and scaled. There are five votes for each of the 36,000 files and these were averaged to obtain a single MOS for each file.

4.3.2 Speech Quality for Tencent Dataset

The Tencent dataset [28] contains speech with simulated online conferencing impairments. Impairments include reverberation, background noise, codecs, packet loss and concealment, filtering, clipping, and combinations of these. Crowdsourced speech-quality ratings were collected and after data cleaning each of the 11,563 files had 20 or more ratings, which were averaged to a single MOS value for each file. We augmented the data by creating a second version of each speech file by trimming 2.3 ms from the front of the file. This is inaudible and allowed us to use the same MOS value for both versions. We accounted for the augmentation so that the unseen testing portion was indeed actually unseen.

4.3.3 Speech Naturalness for Voice Conversion Challenge Dataset

The 2018 Voice Conversion Challenge (VCC2018) evaluated numerous algorithms that convert a speech recording from one (source) talker to sound as if it had been produced by a different (target) talker. One aspect of the evaluation was crowdsourced subjective ratings of how natural the results sound (1 = "completely unnatural" and 5 = "completely natural"). The various combinations of source talker, target talker, utterance, and algorithm produced 20,580 files and all but 15 of these have 4 ratings per file [29]. (Fourteen files have three ratings and one file has two ratings.) The ratings were averaged to yield a single MOS value for each file. We augmented the data as described in 4.3.2.

4.3.4 Speech Quality for NISQA Dataset

We combined the two simulation portions (NISQA_TRAIN_SIM and NISQA_VAL_SIM) of the NISQA database [30] to obtain 12,500 speech files impaired by various types of additive noise, clipping, filtering, the modulated noise reference unit, a range of speech codecs, packet loss conditions, and combinations of these. Crowdsourced ratings of five different dimensions accompany these files, and we used the overall speech quality ratings. Between 3 and 10 overall speech quality ratings are available for each file with an average of 5.2 and a median of 5 ratings per file. The ratings were averaged to yield a single MOS value for each file. We augmented the data as described in 4.3.2.

4.3.5 Speech Quality for PSTN Dataset

The PSTN database [31] is unique in that it holds narrowband speech ($fs = 8$ kHz) consistent with the transmission bandwidth of the legacy public switched telephone network (PSTN). It was created by placing automated calls to pass speech recordings through 80 different PSTN networks (in more than 50 different countries) and then on to one of several VoIP networks where it was then recorded. The database contains 58,709 files. For 69% of these files, noise recordings were mixed with speech recordings before transmission to simulate less desirable acoustic environments. Crowdsourced speech-quality ratings were collected. After data cleaning, each file had between 1 and 10 ratings with an average of 4.6 and a median of 4 ratings per file. The ratings were averaged to yield a single MOS value for each file.

5 TIME-VARYING SPEECH QUALITY EXAMPLE

We have established that FMS can extract useful speech quality and speech naturalness information from entire speech files and that this information can be used to produce estimates of quality and naturalness. These estimates agree with human perception, but not quite as closely as estimates that are produced by the frame-based time-averaged MS. Now we demonstrate that FMS can capture time-varying speech-quality information that is not available after averaging over frame-based MS representations.

We extracted 6,000 speech files ($f_s = 16$ kHz) from the LibriSpeech "dev clean" and "train clean 100" databases [33], and trimmed them to ten seconds. For each file, we randomly selected one of five environmental noise types (bus, car, coffee shop, party, street) and a random segment of that noise type. We used the selected noise segment to create four versions of the speech file: low noise (15 dB SNR), high noise (5 dB SNR), decreasing noise (5 dB SNR for the first 5 seconds, then 15 dB SNR for the final 5 seconds), and increasing noise (15 dB SNR for the first 5 seconds, then 5 dB SNR for the final 5 seconds). In the decreasing noise and increasing noise versions, the noise level ramps down or up over a period of 100 ms. The time of the start of the transition includes a small random component so that the exact start times are uniformly distributed between 4.9 and 5.1 seconds. Note that the first two versions have fixed speech quality and the second two have time-varying speech quality. These 4 versions produce 24,000 speech files total. These four versions allow us to construct a four-way audio impairment classification problem and use modulation spectra and small networks to solve that problem.

For each file we compute the FMS and the frame-based MS as described in 3 and 4. We then create three FMS feature vectors $\boldsymbol{x_0}$: The elements of the magnitude-only feature vector are

$$x_{32m+i} = \log_{10}(\Psi^M_{i,m}), \;\; i = 0 \text{ to } 31, \;\; m = 0 \text{ to } 10, \tag{8}$$

elements of the phase-only feature vector are

$$x_{32m+i} = \log_{10}(\Psi^P_{i,m}), \;\; i = 0 \text{ to } 31, \;\; m = 0 \text{ to } 10, \tag{9}$$

and elements of the magnitude with phase feature vector are

$$\begin{aligned} x_{32m+i} &= \log_{10}(\Psi^M_{i,m}), \;\; i = 0 \text{ to } 31, \;\; m = 0 \text{ to } 10, \\ x_{352+32m+i} &= \log_{10}(\Psi^P_{i,m}), \;\; i = 0 \text{ to } 31, \;\; m = 0 \text{ to } 10. \end{aligned} \tag{10}$$

Similarly, we create the same three options for the frame-based MS: magnitude only,

$$x_{32m+i} = \log_{10}\left(\frac{1}{N_{\text{frm}}}\sum_{f=0}^{N_{\text{frm}}-1}\Psi^M_{i,m,f}\right), \;\; i = 0 \text{ to } 31, \;\; m = 0 \text{ to } 10, \tag{11}$$

phase only,

$$x_{32m+i} = \log_{10}\left(\frac{1}{N_{\text{frm}}}\sum_{f=0}^{N_{\text{frm}}-1}\Psi^P_{i,m,f}\right), \;\; i = 0 \text{ to } 31, \;\; m = 0 \text{ to } 10, \tag{12}$$

and magnitude with phase,

$$x_{32m+i} = \log_{10}\left(\frac{1}{N_{\text{frm}}}\sum_{f=0}^{N_{\text{frm}}-1}\Psi^{M}_{i,m,f}\right), \quad i = 0 \text{ to } 31, \quad m = 0 \text{ to } 10,$$

$$x_{352+32m+i} = \log_{10}\left(\frac{1}{N_{\text{frm}}}\sum_{f=0}^{N_{\text{frm}}-1}\Psi^{P}_{i,m,f}\right), \quad i = 0 \text{ to } 31, \quad m = 0 \text{ to } 10. \tag{13}$$

This gives six types of feature vectors that we can use with networks to solve the four-way classification problem. The networks are as defined in (4) with $N_0 = 352$ (magnitude only or phase only) or $N_0 = 704$ (magnitude with phase). We have $N_1 = N_2 = N_3 = 256$, as before. Finally, $N_4 = 4$, since each network has four outputs. Each output indicates the probability that the input audio signal belongs to one of the four classes (low noise, high noise, decreasing noise, or increasing noise). We find the output with the greatest value and use it as a hard classification. This means the network can assign each input audio file to a single class at inference.

For each of the six types of feature vectors, we used 85% of the data to train a network and 5% to validate it. We use the remaining 10% of the files (2400 files, 6.7 hours) to test the classification capabilities of each network. More specifically, we consider the confusion matrices generated by recording the distribution of inferred classes for each true class. Table 5a through Table 5f show the six confusion matrices. In the tables, the abbreviations *low*, *high*, *dec.*, and *inc.* indicate the classes of low noise, high noise, decreasing noise, and increasing noise, respectively. Note that the ideal confusion matrix has 1.0 as each diagonal entry and 0.0 elsewhere. Table 6 and Table 7 summarize the mean error rates for the six cases.

These confusion matrices can be viewed as images, where 0.0 is shown as black and 1.0 is shown as white, which can be seen in Figure 3. Table 5 through Table 7 and Figure 3 make several results clear. When magnitude is used alone, both frame-based MS and FMS can identify the low and high classes reliably, but the decreasing and increasing classes cannot be properly identified. When phase is used alone, neither frame-based MS nor FMS works well overall. The frame-based approach works better for the low and high classes, while FMS does better with decreasing and increasing classes. When magnitude and phase are used together, the frame-based MS can reliably identify the low and high classes, but it struggles to classify the time-varying classes ("decreasing" and "increasing"). With FMS, magnitude and phase provide sufficient information to reliably identify all four classes, and the mean error rate is roughly one-third that of the frame-based approach. If we focus in on the two cases of time-varying speech quality we see (Table 7) that using magnitude and phase of the FMS gives an error rate that is less than one-fifth the error rate obtained using the frame-based approach. That is, when the MS represents the full audio file, we have a fixed size without any time averaging, and time varying speech quality information is preserved, rather than destroyed, by averaging.

Given that the frame-based approach includes averaging over frames, it may seem surprising that it can do any better than guessing when identifying the "decreasing" and "increasing" cases. The explanation is that some frames span the transition between noise levels and due to the impressive capabilities of MS and ML, these few frames can provide non-zero information regarding the type of transition and thus raise classification performance somewhat above the guessing level.

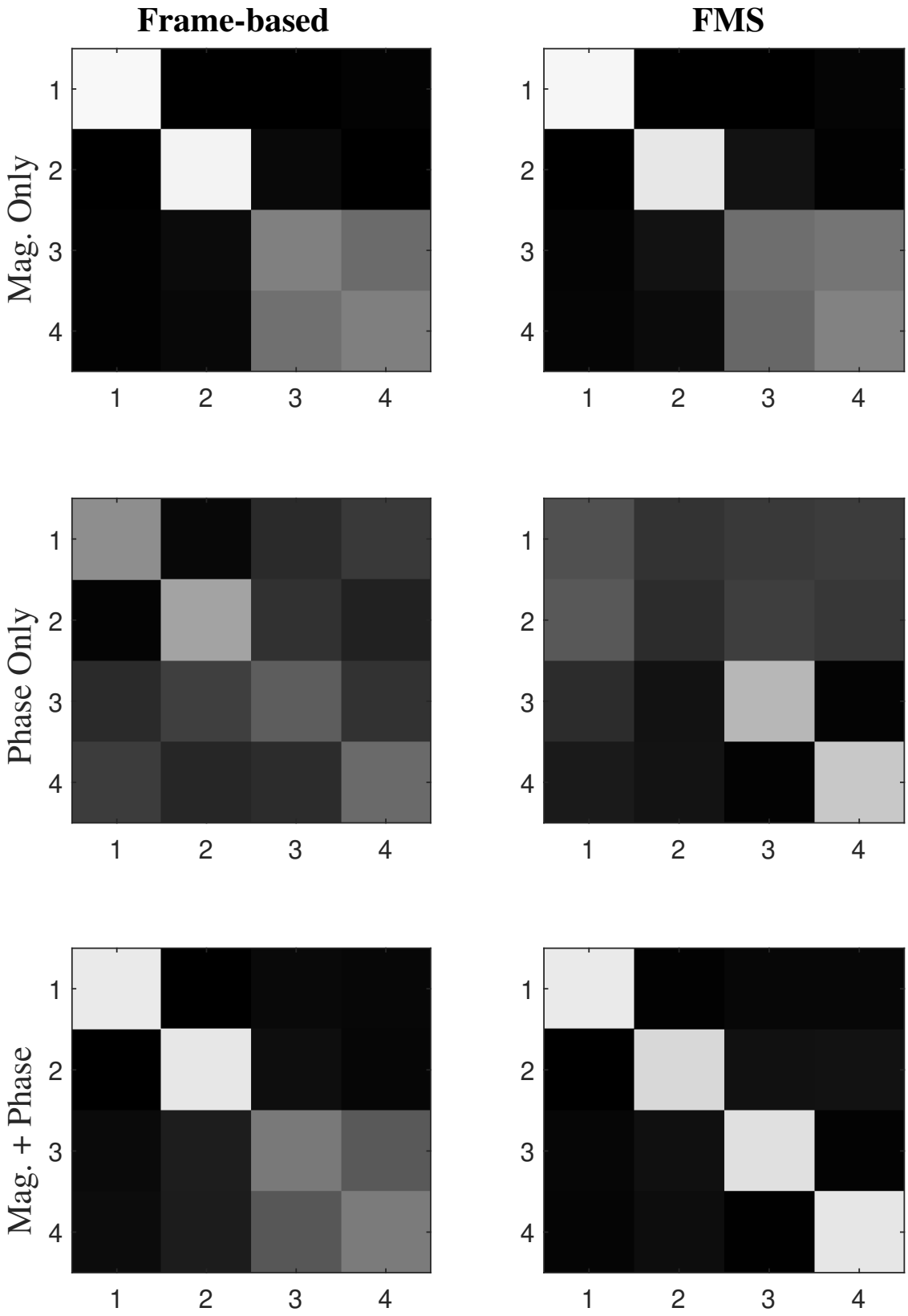

Figure 3. Confusion matrices displayed as images (black indicates 0.0 and white indicates 1.0) Classes 1, 2, 3, and 4 are low noise, high noise, decreasing noise, and increasing noise, respectively.

Table 5. Confusion matrices.

(a) Frame-based MS using magnitude only, mean error rate 0.26.

	Inferred Class			
True Class	Low	High	Dec.	Inc.
Low	0.98	0.00	0.00	0.02
High	0.00	0.96	0.04	0.00
Dec.	0.01	0.05	0.51	0.43
Inc.	0.01	0.04	0.45	0.50

(b) FMS using magnitude only, mean error rate 0.29.

	Inferred Class			
True Class	Low	High	Dec.	Inc.
Low	0.97	0.00	0.00	0.03
High	0.00	0.91	0.08	0.01
Dec.	0.02	0.08	0.44	0.47
Inc.	0.02	0.05	0.41	0.52

(c) Frame-based MS using phase only, mean error rate 0.50.

	Inferred Class			
True Class	Low	High	Dec.	Inc.
Low	0.56	0.04	0.17	0.23
High	0.02	0.64	0.20	0.14
Dec.	0.17	0.25	0.37	0.20
Inc.	0.24	0.16	0.18	0.42

(d) FMS using phase only, mean error rate 0.50.

	Inferred Class			
True Class	Low	High	Dec.	Inc.
Low	0.32	0.21	0.23	0.24
High	0.35	0.18	0.25	0.22
Dec.	0.18	0.08	0.72	0.02
Inc.	0.11	0.08	0.02	0.79

(e) Frame-based MS using magnitude and phase, mean error rate 0.30.

	Inferred Class			
True Class	Low	High	Dec.	Inc.
Low	0.92	0.01	0.04	0.03
High	0.00	0.91	0.06	0.03
Dec.	0.05	0.12	0.48	0.35
Inc.	0.05	0.12	0.35	0.49

(f) FMS using magnitude and phase, mean error rate 0.11.

	Inferred Class			
True Class	Low	High	Dec.	Inc.
Low	0.92	0.01	0.03	0.03
High	0.01	0.85	0.07	0.07
Dec.	0.03	0.07	0.89	0.02
Inc.	0.03	0.06	0.01	0.91

Table 6. Mean error rates for the 4-way classification problem.

	Mag Only	Phase Only	Mag and Phase
frame-based	0.26	0.50	0.30
FMS	0.29	0.50	0.11

Table 7. Mean error rates for identification of decreasing noise and increasing noise cases.

	Mag Only	Phase Only	Mag and Phase
frame-based	0.49	0.60	0.52
FMS	0.52	0.24	0.10

6 DISCUSSION

FMS provide rich, fixed-size, and perceptually consistent representations of speech. Our experiments show that FMS and frame-based MS give similar results when used with simple networks to estimate speech quality and naturalness for five large datasets. A separate experiment shows that, as expected, FMS captures time-varying speech quality information that is mostly lost when the frame-based MS is followed by averaging over frames.

If one elects to use a frame-based MS and not average over frames, more information can be preserved but the size of the representation is increased as well. For the data used here and described in Table 2, representations would be, on average, 212 times larger than FMS. These larger representations have memory implications and naturally increase the computations per file required for training and inference. Further, we found that training networks to use unaveraged frame-based MS inputs has much slower convergence (about three times as many epochs were required) than training for file-based inputs.

From these experiments we conclude that the FMS is a compact, fixed-size representation of an entire, arbitrary-length speech file that is sufficiently descriptive to allow for human-like judgments of telecommunications speech quality and judgments of the naturalness of synthetic speech. One path to better speech quality estimation is to pursue extremely large models, some of which have hundreds of millions of parameters. But it remains important for researchers to peruse smaller and more efficient networks and to push their performance as far as possible. Our work with FMS is one example of that approach. FMS innately encode important temporal information that can be missed when averaging across frame-based representations. FMS can reduce the need for complex temporal processing inside a speech quality estimation network and may allow FMS-based speech quality estimators to use smaller, less complex networks and still achieve performance similar to the best frame-based alternatives.

REFERENCES

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 2020-December, ISSN: 10495258. DOI: 10.48550/arxiv.2006.11477. [Online]. Available: https://arxiv.org/abs/2006.11477v3.

[2] C.-C. Lo, S.-W. Fu, W.-C. Huang, *et al.*, "MOSNet: Deep learning-based objective assessment for voice conversion," in *Proc. Interspeech*, 2019.

[3] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "MBNET: MOS prediction for synthesized speech with mean-bias network," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2021, pp. 391–395. DOI: 10.1109/ICASSP39728.2021.9413877.

[4] R. E. Zezario, S. W. Fu, F. Chen, C. S. Fuh, H. M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Trans.on Audio Speech and Language Processing*, vol. 31, pp. 54–70, 2023, ISSN: 23299304. DOI: 10.1109/TASLP.2022.3205757.

[5] E. Cooper, W. C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2021, pp. 8442–8446, ISBN: 9781665405409. DOI: 10.48550/arxiv.2110.02635. [Online]. Available: https://arxiv.org/abs/2110.02635v3.

[6] W. C. Tseng, C. Y. Huang, W. T. Kao, Y. Y. Lin, and H. Y. Lee, "Utilizing self-supervised representations for MOS prediction," in *Proc. Interspeech*, 2021, ISBN: 9781713836902. DOI: 10.48550/arxiv.2104.03017. [Online]. Available: https://arxiv.org/abs/2104.03017v3.

[7] S. Voran, "A basic experiment on time-varying speech quality," in *Proc. Fourth Intl. Measurement of Speech and Audio Quality in Networks Conference*, Prague, Czech Republic, Jun. 2005, pp. 51–64. [Online]. Available: https://its.ntia.gov/audio.

[8] J. Berger, A. Hellenbart, R. Ullmann, *et al.*, "Estimation of 'quality per call' in modelled telephone conversations," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2008, pp. 4809–4812. DOI: 10.1109/ICASSP.2008.4518733.

[9] D. S. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 821–831, 2005. DOI: 10.1109/TSA.2005.851924.

[10] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010. DOI: 10.1109/TASL.2010.2052247.

[11] J. F. Santos, S. Cosentino, O. Hazrati, P. C. Loizou, and T. H. Falk, "Objective speech intelligibility measurement for cochlear implant users in complex listening environments," en, *Speech Communications*, vol. 55, no. 7-8, pp. 815–824, 2013.

[12] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Proc. 14th Intl. Workshop on Acoustic Signal Enhancement*, 2014, pp. 55–59. DOI: 10.1109/IWAENC.2014.6953337.

[13] D. Kim and A. Tarraf, "ANIQUE+: A new American National Standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Technical Journal*, vol. 12, no. 1, pp. 221–236, Spring 2007, ISSN: 1538-7305. DOI: 10.1002/bltj.20228.

[14] B. Cauchi, J. F. Santos, K. Siedenburg, *et al.*, "Predicting the quality of processed speech by combining modulation-based features and model trees," in *Proc. Twelfth ITG Conference on Speech Communication*, 2016.

[15] M. Hakami and W. B. Kleijn, "Machine learning based non-intrusive quality estimation with an augmented feature set," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2017, pp. 5105–5109. DOI: 10.1109/ICASSP.2017.7953129.

[16] B. Cauchi, K. Siedenburg, J. F. Santos, T. H. Falk, S. Doclo, and S. Goetze, "Non-intrusive speech quality prediction using modulation energies and LSTM-network," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1151–1163, 2019. DOI: 10.1109/TASLP.2019.2912123.

[17] D. O'Shaughnessy, *Speech Communication: Human and Machine*. Addison-Wesley, 1987.

[18] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2019, pp. 7125–7129.

[19] A. A. Catellier and S. D. Voran, "Wideband audio waveform evaluation networks: Efficient, accurate estimation of speech qualities," *IEEE Access*, vol. 11, pp. 125 576–125 592, 2023. DOI: 10.1109/ACCESS.2023.3330640.

[20] X. Dong and D. S. Williamson, "A pyramid recurrent network for predicting crowdsourced speech-quality ratings of real-world signals," in *Proc. Interspeech*, 2020.

[21] D.-S. Kim, "A cue for objective speech quality estimation in temporal envelope representations," *IEEE Signal Processing Letters*, vol. 11, no. 10, pp. 849–852, 2004. DOI: 10.1109/LSP.2004.835466.

[22] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.

[23] O. Ghitza, "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," *The Journal of the Acoustical Society of America*, vol. 110, no. 3, Pt. 1, pp. 1628–1640, 2001.

[24] N. F. Viemeister, "Temporal modulation transfer functions based upon modulation thresholds," *The Journal of the Acoustical Society of America*, vol. 66, no. 5, pp. 1364–1380, 1979.

[25] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction.," *The Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1593–602, 1994, ISSN: 0001-4966.

[26] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2022, pp. 886–890. DOI: 10.1109/ICASSP43922.2022.9746108.

[27] K. E. Hajal, M. Cernak, and P. Mainar, "MOSRA: Joint mean opinion score and room acoustics speech quality assessment," in *Proc. Interspeech*, 2022.

[28] Gaoxiong Yi et al., "ConferencingSpeech 2022 Challenge: Non-intrusive objective speech quality assessment challenge for online conferencing applications," in *Proc. Interspeech*, 2022.

[29] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, *et al.*, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Speaker Odyssey*, 2018.

[30] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *INTERSPEECH*, pp. 2127–2131, 2021.

[31] G. Mittag, R. Cutler, Y. Hosseinkashi, *et al.*, "DNN no-reference PSTN speech quality prediction," in *Proc. Interspeech*, Oct. 2020. DOI: 10.21437/interspeech.2020-2760.

[32] S. Voran and A. Catellier, "A crowdsourced speech intelligibility test that agrees with, has higher repeatability than, lab tests," NTIA, Tech. Rep. TM-17-523, 2017.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

Appendix A: Mel Spectrum Filter Bank

The mel spectrum filter bank used to create the fixed-size modulation spectrum (FMS) accumulates Hertz-scale spectral samples to create a mel-scale spectral representation [17]. The filters have uniform width and spacing on the mel scale and triangular shape on the Hertz scale. There are N_{mel} filters $\theta_{k,i}$, $i = 0$ to $N_{mel} - 1$, $k = 0$ to $N_{Hertz} - 1$. Since the length of the DFT that produces the Hertz-scale spectral representation is N_t (even), it follows that $N_{Hertz} = (N_t/2) + 1$.

The upper frequency limit for the analysis is u_f (in Hz). This value is converted to a mel-scale value $\tilde{u}_f$ (in mel) using the relationship given in [17],

$$\tilde{u}_f = 2595 \log_{10}(1 + u_f/700), \tag{14}$$

and the resulting range is evenly divided:

$$\tilde{\Delta} = \tilde{u}_f/(N_{mel} + 1). \tag{15}$$

Let $\tilde{b}_i = i\tilde{\Delta}$, for $i = 0$ to $N_{mel} + 1$. Then the filter centered at $\tilde{b}_i$ ($i = 1$ to N_{mel}) extends from $\tilde{b}_{i-1}$ to $\tilde{b}_{i+1}$. We use the inverse of (14) to convert the $\tilde{b}_i$ to their Hertz scale equivalents b_i:

$$b_i = 700 \left(10^{(\tilde{b}_i/2595)} - 1 \right). \tag{16}$$

The filter bank values are given by

$$\theta_{k,i} = \begin{cases} \eta_i \frac{f_k - b_i}{b_{i+1} - b_i}, & b_i \le f_k < b_{i+1}, \\ \eta_i \left(1 - \frac{f_k - b_{i+1}}{b_{i+2} - b_{i+1}} \right), & b_{i+1} \le f_k < b_{i+2}, \\ 0, & \text{otherwise}, \end{cases}$$
$$k = 0 \text{ to } N_{Hertz} - 1, \quad i = 0 \text{ to } N_{mel} - 1. \tag{17}$$

The Hertz-scale frequencies f_k used in (17) are calculated from the audio sample rate f_s and the DFT length N_t (even):

$$f_k = k \frac{f_s}{N_t}, \quad k = 0 \text{ to } N_t/2. \tag{18}$$

The limits for k given in (17) and (18) are consistent because $N_{Hertz} = (N_t/2) + 1$. Normalization for the width of each band is accomplished by

$$\eta_i = \frac{1}{b_{i+2} - b_i}, \quad i = 0 \text{ to } N_{mel} - 1. \tag{19}$$

Using (17), we can show that the upper slope of filter $\theta_{\cdot,i-1}$ and the lower slope of filter $\theta_{\cdot,i}$ intersect at

$$\frac{\eta_{i-1} b_{i+1} + \eta_i b_i}{\eta_{i-1} + \eta_i} \text{ Hz}. \tag{20}$$

In the simplified case $\eta_{i-1} = \eta_i$, the intersection given in (20) becomes the midpoint of b_i and b_{i+1}. We can also use (17) to show that the filter centered at b_i has a peak value of η_i and at the simplified intersection locations $(b_i + b_{i-1})/2$ and $(b_i + b_{i+1})/2$ it has value $\eta_i/2$. Thus, at the simplified intersection location, the value of the filter relative to its peak value is $1/2$ or minus 6.0 dB.

Appendix B: Modulation Spectrum Rectangular Filter Bank

The rectangular modulation spectrum filter bank contains N_{mod} filters and each performs a different aggregation of N linear-Hertz-scale spectral samples. The result is a logarithmic-Hertz-scale spectral representation with N_{mod} samples. When viewed on a logarithmic frequency scale, the filters have uniform spacing and they place the same weight on each spectral sample (i.e., they have rectangular shape). The N_{mod} filters can be described as elements in a matrix $\Phi_{k,m}$, $m = 0$ to $N_{mod} - 1$, $k = 0$ to $N - 1$.

The first filter simply selects the linear-Hertz-scale spectral sample at DC. The remaining $N_{mod} - 1$ filters collect multiple linear-Hertz-scale spectral samples to produce a log-scale output. The first of these is nominally located at 0.25 Hz. The last is nominally at 128 Hz. Thus, it is required that $3 \leq N_{mod}$.

Since the number of per-frame results used in the modulation spectrum DFT is N_f, the number of resulting unique spectral values is $\lfloor N_f/2 \rfloor + 1$. This means that the number of linear-Hertz-scale spectral values input to the filter bank is $N = \lfloor N_f/2 \rfloor + 1$.

The modulation spectrum filter bank is constructed as follows. The range from 0.25 to 128 Hz is evenly divided on the log scale:

$$\bar{\Delta} = \frac{\log_2(128) - \log_2(0.25)}{N_{mod} - 2}. \tag{21}$$

The log-scale center frequencies are given by

$$\bar{b}_m = \log_2(0.25) + m\bar{\Delta}, \quad m = 0 \text{ to } N_{mod} - 2. \tag{22}$$

These $N_{mod} - 1$ center frequencies are used to define N_{mod} thresholds, δ_i, given by

$$\delta_i = \begin{cases} -\infty, & i = 0, \\ \frac{1}{2}(\bar{b}_{i-1} + \bar{b}_i), & i = 1 \text{ to } N_{mod} - 2, \\ +\infty, & i = N_{mod} - 1. \end{cases} \tag{23}$$

The frequencies of the DFT bins can be calculated from the audio sample rate, f_s, the frame stride, N_s (number of samples of advance when forming the next frame), and the number of frames used in the modulation spectrum DFT, N_f:

$$\Delta_h = \frac{f_s}{N_s N_f}. \tag{24}$$

The log frequencies of the DFT bins are given by

$$\bar{f}_k = \log_2(k\Delta_h), \quad k = 0 \text{ to } N - 1, \tag{25}$$

with $\log_2(0)$ defined to be $-\infty$.

We can now define the filter bank values:

$$\begin{aligned} &\Phi_{0,0} = 1, \\ &\Phi_{k,0} = 0, \;\; k = 1 \text{ to } N-1, \end{aligned} \tag{26}$$

and for $m = 1$ to $N_{mod} - 1$, $k = 0$ to $N-1$,

$$\Phi_{k,m} = \begin{cases} \nu_m, & \delta_{m-1} < \bar{f}_k \leq \delta_m, \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

The weights ν_m serve to normalize each filter by the number of DFT samples it spans:

$$\nu_m = \frac{1}{\left|\left\{k \in \mathbb{Z} : \delta_{m-1} < \bar{f}_k \leq \delta_m\right\}\right|}, \quad m = 1 \text{ to } N_{mod} - 1, \tag{28}$$

where $|\cdot|$ indicates set cardinality in this context.

Appendix C: Modulation Spectrum Triangular Filter Bank

The triangular modulation spectrum filter bank accumulates N linear-Hertz-scale spectral samples to create a logarithmic-Hertz-scale spectral representation with N_{mod} samples. When viewed on a logarithmic frequency scale, the filters have uniform spacing and identical triangular shape. Following the successful precedent established in [9], the first filter is centered at 4 Hz and the last filter is centered at 128 Hz, so it is required that $2 \leq N_{mod}$. The filter bank contains N_{mod} filters and they can be described as elements of the matrix $\Phi_{k,m}$, $m = 0$ to $N_{mod} - 1$, $k = 0$ to $N - 1$. Since the number of per-frame results used in the modulation spectrum DFT is N_f, the number of resulting unique spectral values is $\lfloor N_f/2 \rfloor + 1$. In other words, the number of linear-Hertz-scale spectral values used as input to the filter bank is $N = \lfloor N_f/2 \rfloor + 1$.

The modulation spectrum filter bank is constructed as follows. The range from 4 to 128 Hz is evenly divided on the log scale:

$$\bar{\Delta} = \frac{\log_2(128) - \log_2(4)}{N_{mod} - 1}. \tag{29}$$

The initial log-scale center frequencies are given by

$$\bar{b}_m = \log_2(4) + m\bar{\Delta}, \;\; m = 0 \text{ to } N_{mod} - 1. \tag{30}$$

To maximize consistency across different file lengths, we move each of these initial filter center frequencies to match the nearest DFT bin. The spacing of the DFT bins can be calculated from the audio sample rate f_s, the frame stride N_s (number of samples of advance when forming the next frame) and the number of frames used in the modulation spectrum DFT N_f:

$$\Delta_h = \frac{f_s}{N_s N_f}, \tag{31}$$

and the log frequencies of the DFT bins are given by

$$\bar{f}_k = \log_2(k\Delta_h), \quad k = 0 \text{ to } N - 1, \tag{32}$$

with $\log_2(0)$ defined to be $-\infty$. Moving $\bar{b}_m$ to match the nearest value of $\bar{f}_k$ is achieved by

$$\bar{b}'_m = \log_2\left(\left[\frac{2^{\bar{b}_m}}{\Delta_h}\right]\Delta_h\right), \;\; m = 0 \text{ to } N_{mod} - 1, \tag{33}$$

where $[\cdot]$ indicates rounding to the nearest integer. Due to the logarithmic scale, these adjustments are very small near the top of the frequency scale (128 Hz), but larger near the bottom of the scale (4 Hz). The filter half-widths are given by

$$\bar{\delta} = \bar{\Delta}/(2 - \sqrt{2}), \tag{34}$$

and these values are selected to make unweighted adjacent filters intersect at minus 3 dB.

The filter bank values are given by

$$\hat{\Phi}_{k,m} = \begin{cases} \nu_m \frac{\bar{f}_k - (\bar{b}'_m - \bar{\delta})}{\bar{\delta}}, & \bar{b}'_m - \bar{\delta} \le \bar{f}_k < \bar{b}'_m, \\ \nu_m \left(1 - \frac{\bar{f}_k - \bar{b}'_m}{\bar{\delta}}\right), & \bar{b}'_m \le \bar{f}_k < \bar{b}'_m + \bar{\delta}, \\ 0, & \text{otherwise}, \end{cases}$$
$$k = 0 \text{ to } N-1, \qquad m = 0 \text{ to } N_{mod} - 1. \tag{35}$$

The weights, ν_m, serve to normalize each filter by the number of DFT samples it spans:

$$\nu_m = \frac{1}{\left|\left\{k \in \mathbb{Z} : -\bar{\delta} \le \bar{f}_k - \bar{b}'_m < \bar{\delta}\right\}\right|}, \quad m = 0 \text{ to } N_{mod} - 1, \tag{36}$$

where $|\cdot|$ indicates set cardinality in this context.

Using (35), we can show that the upper slope of filter $\hat{\Phi}_{\cdot,i}$ and the lower slope of filter $\hat{\Phi}_{\cdot,i+1}$ intersect at

$$\frac{\nu_i(\bar{b}'_i + \bar{\delta}) + \nu_{i+1}(\bar{b}'_{i+1} - \bar{\delta})}{\nu_i + \nu_{i+1}} \quad \log_2(\text{Hz}). \tag{37}$$

In the simplified case of $\nu_i = \nu_{i+1}$, (37) becomes the midpoint of $\bar{b}'_i$ and $\bar{b}'_{i+1}$.

We can also use (35) to show that the filter centered at $\bar{b}'_i$ has a peak value of ν_i and at the simplified intersection locations $(\bar{b}'_i + \bar{b}'_{i+1})/2$ it has value $\nu_i/\sqrt{2}$. Thus, at the simplified intersection location, the value of the filter relative to the peak value is $1/\sqrt{2}$ or minus 3.0 dB.

Form NTIA-29
(6/04)

U.S. Department of Commerce
National Telecommunications and Information Administration

BIBILIOGRAPHIC DATA SHEET

1. Publication Number	2. Government Accession Number	3. Recipient's Accession Number
NTIA TM-24-574		

4. Title and Subtitle

A Powerful, Fixed-Size Modulation Spectrum Representation for Perceptually Consistent Speech Evaluation

5. Publication Date

September 2024

6. Performing Organization Code

NTIA/ITS.P

7. Author(s)

Stephen D. Voran and Jaden Pieper

9. Project/Task/Work Unit No.

3142012-300

8. Performing Organization Name and Address

National Telecommunications and Information Administration
Institute for Telecommunication Sciences, 325 Broadway, Boulder, CO 80305

10. Contract/Grant Number

11. Sponsoring Organization Name and Address

National Telecommunications and Information Administration
Herbert C. Hoover Bldg. 14th & Constitution Ave., NW,
Washington, DC 20230

12. Type of Report and Period Covered

14. Supplementary Notes

15. ABSTRACT *(A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)*

We develop the wideband fixed-size modulation spectra (FMS) and show that they contain the necessary information to perform perceptually consistent evaluation of speech. We compare FMS with the already established frame-based modula-tion spectra as representations for estimating speech quality and speech natural-ness. We feed the two representations into equally sized, relatively small, fully connected networks for five proof-of-concept experiments and find that the two representations perform similarly when estimating speech quality and speech nat-uralness. But the FMS representation captures an entire speech file into a fixed size representation, which means that additional temporal processing is neither needed nor possible. This is in contrast to the frame-based representations, where addi-tional processing can either destroy information (e.g., averaging over time) or lead to more complex and difficult to train networks.

We also demonstrate that when speech quality changes within a speech file, FMS has another distinct advantage which is to be able to efficiently and reliably iden-tify different situations in a way that is not well-addressed by the frame-based approach. Our experiments include more than 274 hours of speech in two languages. This speech is contained in 139,000 speech files and there is a subjective score for each file. File lengths range from 0.6 to 28 seconds and five different sample rates are present. Supporting software is available at https://github.com/NTIA/fms.

16. Key Words *(Alphabetical order, separated by semicolons)*

Auditory perception, fixed-size modulation spectra (FMS), machine learning, mel spectrum, modulation spectrum, speech naturalness, speech quality, time-varying speech quality

17. Availability Statement	18. Security Class. *(This report)*	20. Number of Pages
☑ Unlimited	Unclassified	29
☐ For Official Distribution	**19. Security Class.** *(This page)*	**21. Price**
	Unclassified	N/A

NTIA FORMAL PUBLICATION SERIES

NTIA MONOGRAPH (MG)

A scholarly, professionally oriented publication dealing with state-of-the-art research or an authoritative treatment of a broad area. Expected to have long-lasting value.

NTIA SPECIAL PUBLICATION (SP)

Conference proceedings, bibliographies, selected speeches, course and instructional materials, directories, and major studies mandated by Congress.

NTIA REPORT (TR)

Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities.

JOINT NTIA/OTHER-AGENCY REPORT (JR)

Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities.

NTIA SOFTWARE & DATA PRODUCTS (SD)

Software such as programs, test data, and sound/video files. This series can be used to transfer technology to U.S. industry.

NTIA HANDBOOK (HB)

Information pertaining to technical procedures, reference and data guides, and formal user's manuals that are expected to be pertinent for a long time.

NTIA TECHNICAL MEMORANDUM (TM)

Technical information typically of less breadth than an NTIA Report. The series includes data, preliminary project results, and information for a specific, limited audience.

For information about NTIA publications, contact the NTIA/ITS Technical Publications Office at 325 Broadway, Boulder, CO, 80305 Tel. (303) 497-3572 or e-mail ITSinfo@ntia.gov.