MATHEMATICAL ANALYSIS OF SPEECH FEATURE EXTRACTION FROM RAW WAVEFORMS: FROM SPECTRAL REPRESENTATIONS TO WAV2VEC 2.0

Rohmonqulov Muhammadyusuf Egamberdi o'g'li
Mirzo Ulug'bek nomidagi O'zbekiston Milliy universiteti Jizzax filiali
muhammadyusufrohmonqulov@gmail.com

Abstract. This analytical study formalizes the transformation of a sampled speech waveform into machine-usable acoustic features. The classical path is expressed through short-time framing, Hamming windowing, Fourier analysis, Mel filter banks and MFCCs, while the learned path is examined using wav2vec 2.0. A formula-based derivation shows that the seven-layer wav2vec 2.0 convolutional encoder has an overall stride of 320 samples and a receptive field of 400 samples; at 16 kHz these correspond to 20 ms feature spacing and 25 ms acoustic context. Thus, classical and learned front ends operate on comparable local time scales but differ in whether the transformation is fixed by signal-processing assumptions or learned from unlabeled speech.

Keywords: raw waveform, speech feature extraction, MFCC, wav2vec 2.0, self-supervised learning.

1. INTRODUCTION

After analog-to-digital conversion, speech is represented as a discrete sequence x[n], not as an acoustic pressure wave. Feature extraction converts this high-dimensional waveform into representations that retain phonetic information while reducing irrelevant variability. Classical speech systems construct such representations using short-time spectral analysis and perceptual transforms; modern self-supervised systems learn the transformation from raw audio. MFCC remains a standard interpretable descriptor [1], [2], whereas wav2vec 2.0 is a representative learned front end based on temporal convolution and Transformer context modeling [3], [4].

The aim of this work is to analyze, mathematically and step by step, how useful speech properties are obtained from the raw waveform and to show where the classical MFCC pipeline and wav2vec 2.0 are structurally similar or different. The study is analytical rather than an accuracy benchmark.

2. MATERIALS AND METHODS

Let $x_a(t)$ be the continuous speech signal and F_s the sampling frequency. After sampling and quantization, $x[n] = Q\{x_a(n/F_s)\}$, $n = 0,\ldots,N-1$. For $F_s = 16$ kHz, one second contains 16,000 samples. A normalized waveform may be written as

$$\tilde{x}[n] = (x[n] - \mu_x) / \sqrt{(\sigma_x^2 + \varepsilon)}, \quad \mu_x = (1/N)\Sigma_n x[n]. \qquad (1)$$

Because speech is approximately stationary only over short intervals, the waveform is divided into overlapping frames. With frame duration T_f and hop T_h, $N_f = \text{round}(T_f F_s)$ and $H = \text{round}(T_h F_s)$. For 25 ms and 10 ms at 16 kHz, $N_f = 400$ and $H = 160$. The m-th frame is commonly tapered by a Hamming window:

$$x_m[n] = \tilde{x}[n+mH]w[n], \quad w[n] = 0.54 - 0.46\cos(2\pi n/(N_f-1)). \tag{2}$$

The window reduces spectral leakage caused by abrupt frame boundaries. Each frame is then transformed from time to frequency by the discrete Fourier transform; the FFT is its efficient computational implementation:

$$X_m[k] = \Sigma_{n=0}^{Nf-1} x_m[n]e^{\wedge}(-j2\pi kn/N_FFT), \quad P_m[k] = |X_m[k]|^2/N_FFT. \tag{3}$$

The power spectrum $P_m[k]$ is integrated by triangular filters whose center frequencies are spaced on the Mel scale, $Mel(f)=2595\log_{10}(1+f/700)$. For filter $H_r[k]$,

$$E_m[r] = \Sigma_k P_m[k]H_r[k], \quad L_m[r] = \ln(E_m[r]+\varepsilon). \tag{4}$$

A discrete cosine transform then compacts the smooth log-spectrum into cepstral coefficients:

$$c_m[p] = \Sigma_{r=1}^{M} L_m[r]\cos[\pi p(r-1/2)/M], \quad p=0,\ldots,P-1. \tag{5}$$

Retaining about 12–13 low-order coefficients gives a compact description of the vocal-tract spectral envelope. Therefore, MFCC explicitly encodes assumptions about short-time stationarity, spectral magnitude, auditory frequency resolution, logarithmic compression and cepstral compaction [1], [2].

wav2vec 2.0 replaces these fixed spectral operations with a trainable mapping $z=f_\theta(x)$. A simplified temporal convolution layer can be written as

$$z_l[t] = \mathrm{GELU}(\mathrm{Norm}(\Sigma_i W_l[i]z_{l-1}[t\cdot s_l+i]+b_l)), \quad z_0=x. \tag{6}$$

Its BASE feature encoder contains seven 1-D convolutional blocks with 512 channels, kernel widths $k=(10,3,3,3,3,2,2)$ and strides $s=(5,2,2,2,2,2,2)$ [3]. Hence the spacing between adjacent latent vectors is

$$S = \prod_{l=1}^{7} s_l = 5\cdot 2^6 = 320 \text{ samples} = 20 \text{ ms at 16 kHz}. \tag{7}$$

The receptive field follows $R_0=1$, $J_0=1$, $R_l=R_{l-1}+(k_l-1)J_{l-1}$ and $J_l=J_{l-1}s_l$. Successive values are $R=(10,20,40,80,160,240,400)$, so the final latent vector observes 400 waveform samples, or 25 ms. Thus wav2vec 2.0 learns a local acoustic analysis scale close to the conventional 25 ms frame without explicitly defining a frame or Fourier transform.

The latent sequence is passed to a Transformer context network. For one attention head,

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(QK^T/\sqrt{d_k})V. \tag{8}$$

During self-supervised pre-training, masked latent positions are predicted through a contrastive objective over quantized targets. This forces the representation to preserve acoustic distinctions that are useful across surrounding speech context rather than only within a fixed spectrum [3], [4].

3. RESULTS AND DISCUSSION

The analysis produces a direct numerical correspondence: both a conventional 25 ms frame at 16 kHz and one wav2vec 2.0 encoder receptive field contain 400 input samples. Their representations, however, are generated differently. MFCC applies a

predetermined chain—windowing, Fourier magnitude, Mel integration, logarithm and DCT—so each coefficient has a relatively interpretable spectral origin. wav2vec 2.0 learns convolution kernels from data and then uses attention to add longer-range context. It therefore avoids committing in advance to fixed frequency bands or a fixed cepstral basis.

This distinction is important for pronunciation analysis. MFCCs are compact and transparent for measuring spectral-envelope changes, but information removed by phase elimination, Mel integration and low-order DCT cannot be reconstructed. Learned representations can retain a broader combination of amplitude, periodicity, spectral shape and temporal context, although individual dimensions are harder to interpret; analyses of wav2vec encoders have shown that information related to fundamental frequency, formants and amplitude is present in their latent space [5]. A sound experimental design can therefore use MFCC as an interpretable baseline and wav2vec 2.0 as a learned representation, comparing both at matched temporal resolution.

4. CONCLUSION

Raw-waveform feature extraction is fundamentally a mathematical transformation from sample amplitudes to phonetic representations. MFCC constructs this transformation explicitly through short-time spectral and perceptual operations. wav2vec 2.0 learns the corresponding local transformation from data: at 16 kHz its encoder advances by 320 samples and observes 400 samples per latent vector, after which Transformer attention introduces broader context. The methodological shift is therefore from fixed human-designed transforms to parameterized transforms optimized by self-supervised learning, not from signal processing to the absence of signal processing.

REFERENCES

[1] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoust., Speech, Signal Process., vol. 28, no. 4, pp. 357–366, 1980, doi: 10.1109/TASSP.1980.1163420.

[2] Z. Kh. Abdul and A. K. Al-Talabani, “Mel Frequency Cepstral Coefficient and its applications: A review,” IEEE Access, vol. 10, pp. 122136–122158, 2022, doi: 10.1109/ACCESS.2022.3223444.

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Adv. Neural Inf. Process. Syst., vol. 33, pp. 12449–12460, 2020.

[4] A. Vaswani et al., “Attention is all you need,” Adv. Neural Inf. Process. Syst., vol. 30, pp. 5998–6008, 2017.

[5] K. Choi and E. J. Yeo, “Opening the black box of wav2vec feature encoder,” arXiv:2210.15386, 2022.

IOT TEXNOLOGIYALARI YORDAMIDA RAQAMLASHTIRILGAN ENERGIYA TA’MINOT OBYEKTLARIDA AXBOROT XAVFSIZLIGINI TA'MINLASH MODEL VA ALGORITMLARI.