Voiced/Unvoiced Classification in Speech: Comparing MFCC and Mel Spectrogram CNNs

Nassim Arifette [1]

Abstract

This study investigates the classification of voiced and unvoiced segments in speech signals using both traditional and modern machine learning approaches. We examine conventional methods such as Mel-Frequency Cepstral Coefficients (MFCC) and Mel spectrogram analysis alongside contemporary deep learning techniques. Our research focuses on the application of convolutional neural networks to datasets containing speech samples from three speakers, evaluated using frame-level voiced/unvoiced classification with 7-fold cross-validation. We compare CNN performance on MFCC features versus Mel spectrograms at 32, 64, and 128 bands. Results show MFCC features achieve about 85% validation accuracy, while Mel spectrograms achieve about 81–82% across band configurations. Accurate voiced/unvoiced classification is fundamental to speech processing applications, as it serves as a prerequisite for fundamental frequency estimation, where voiced segments typically exhibit greater harmonic content and periodicity than unvoiced segments.

1. Introduction

The classification of voiced and unvoiced segments in speech signals plays a crucial role in diverse speech processing applications, including speech recognition, speaker verification, and speech synthesis (Hillenbrand et al., 1994). Voiced/unvoiced classification presents significant challenges due to factors such as background noise, breathy or creaky voice quality, fricative sounds with high-frequency energy, and ambiguous boundaries between speech segments. Traditional approaches to voiced/unvoiced classification typically rely on manually engineered features and heuristic-based algorithms (Jouvet & Laprie, 2017). However, these methods may fail to capture complex patterns and variations inherent in speech signals, necessitating the adoption of advanced techniques capable of automatically learning discriminative features from raw speech waveforms to improve classification accuracy.

In recent years, machine learning methodologies, particularly deep learning, have achieved remarkable success in audio signal processing (Huang et al., 2014). These approaches have the potential to revolutionize voiced/unvoiced classification by enabling automatic extraction of high-level representations from raw speech data. Deep neural networks, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can learn hierarchical features that capture complex relationships within speech signals, thereby improving classification performance.

The choice between MFCC and Mel spectrogram representations is particularly interesting in this context. MFCCs decorrelate log-Mel energies and emphasize spectral envelope characteristics that often align with voicing cues, while Mel spectrograms preserve local harmonic structure that may be better captured by CNNs with sufficient data and context.

The objective of this investigation is to leverage machine learning, specifically deep learning, for voiced/unvoiced classification in speech signals. We employ a CNN architecture for accurate discrimination between voiced and unvoiced audio segments. Through training on labeled audio samples from three speakers, our model learns to identify discriminative patterns associated with each category.

In the following sections, we outline our methodology, present experimental results, and discuss the implications of our research. By combining the power of deep learning with the challenges of voiced/unvoiced classification, we aim to contribute to the development of more sophisticated and accurate speech processing systems.

2. Audio Representation: Mel Spectrogram and MFCCs

In speech signal processing, selecting an appropriate audio representation is critical for accurate analysis and classification (Zhou et al., 2011). Two widely adopted representations

[1] Université Paris-Saclay, France. Correspondence to: Nassim Arifette <nassim.arifette@universite-paris-saclay.fr>.

are the Mel spectrogram and Mel-Frequency Cepstral Coefficients (MFCCs) (Sato & Obuchi, 2007; Ayvaz et al., 2022). These representations capture important spectral characteristics of audio signals and have been extensively utilized in various applications such as speech recognition, speaker identification, and music analysis.

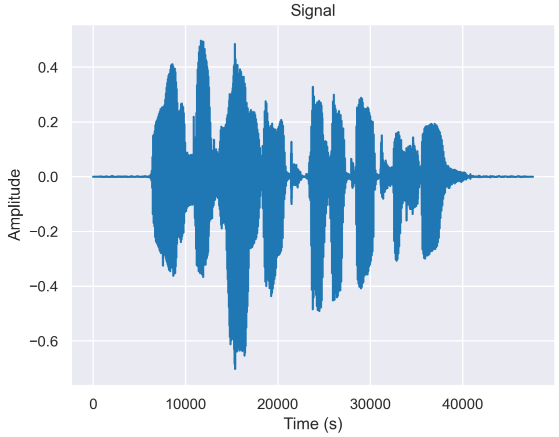

Figure 1. Audio signal waveform from one speaker (25 ms analysis window, 10 ms hop size, 16 kHz sampling rate).

2.1. Mel Spectrogram

The Mel spectrogram is obtained by applying a Mel filterbank to the short-time Fourier transform (STFT) magnitude spectrum of an audio signal. The Mel filterbank divides the frequency spectrum into perceptually relevant frequency bands, approximating the non-linear characteristics of human auditory perception. This transformation provides a spectrogram where the frequency axis is scaled according to the Mel scale, resulting in a representation that better aligns with human auditory perception.

The Mel spectrogram calculation begins with the STFT using a 25 ms Hann window with 10 ms hop size:

$$X(k,n) = \sum_{t=0}^{N-1} x(t+nH)w(t)e^{-j2\pi kt/N}$$

where $x(t)$ is the audio signal sampled at 16 kHz, $w(t)$ is the Hann window function, $H = 160$ samples (10 ms hop), $N = 400$ samples (25 ms window), and $X(k,n)$ is the STFT coefficient at frequency bin k and time frame n.

We then apply the Mel filterbank with M filters spanning 0–8 kHz (Slaney-style triangular filters (Slaney, 1998)) to obtain the Mel spectrogram:

$$S(m,n) = \sum_{k=0}^{N/2} |X(k,n)|^2 \, H_m(k)$$

where $H_m(k)$ is the m-th Mel filter response at frequency bin k, $S(m,n)$ is the Mel spectrogram power at Mel band m and time frame n, and $M \in \{32, 64, 128\}$ depending on the experimental configuration.

Finally, we apply logarithmic compression:

$$S_{\log}(m,n) = 10\log_{10}(S(m,n)+\epsilon)$$

where $\epsilon = 10^{-10}$ is a small constant to avoid taking the logarithm of zero, resulting in features expressed in dB. We use $10\log_{10}$ compression; using the natural logarithm would change only a constant scale and not the relative information content.

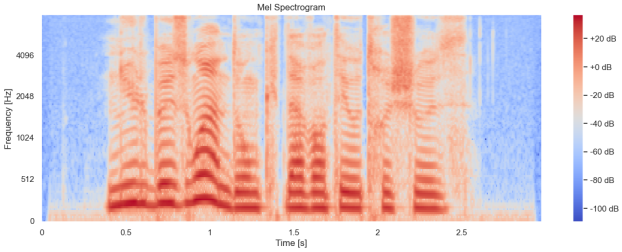

Figure 2. Mel spectrogram corresponding to the audio signal in Figure 1 (64 Mel bands, 0–8 kHz frequency range).

2.2. Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCCs) are obtained by applying the Discrete Cosine Transform Type-II (DCT-II) to the logarithmic Mel spectrogram:

$$C(k,n) = \sqrt{\frac{2}{M}} \sum_{m=0}^{M-1} S_{\log}(m,n) \cdot \cos\left[\frac{\pi k(2m+1)}{2M}\right]$$

where $C(k,n)$ is the k-th MFCC coefficient at time frame n, $M = 26$ is the number of Mel filters for MFCC computation, and the normalization factor $\sqrt{2/M}$ applies for $k > 0$ (for $k = 0$, the factor is $\sqrt{1/M}$).

In our implementation, we retain 20 MFCC coefficients, apply a liftering coefficient of 22, and do not include delta (Δ) or delta-delta ($\Delta\Delta$) coefficients in the baseline experiments. The MFCC computation uses the same windowing parameters as described above (25 ms Hann windows, 10 ms hop). Per-utterance cepstral mean and variance normalization (CMVN) is applied to improve robustness. We exclude C_0 to reduce sensitivity to overall loudness and channel effects, improving robustness across recordings.

Unlike representations that emphasize low-frequency components, the MFCC transformation separates source-filter components of speech. The lower cepstral coefficients encapsulate slow spectral variations that reflect vocal tract filtering characteristics, while higher cepstral coefficients capture finer spectral details related to source excitation, noise levels, and other factors.

2.3. Fundamental Frequency (F0) in Voiced and Unvoiced Speech

The classification of speech sounds into voiced and unvoiced categories is a fundamental aspect of phonetics (Fant, 1971). Voiced sounds are characterized by periodic vibration of the vocal folds, generating regular pressure variations that result in sound with a measurable fundamental frequency (F0). Voiced sounds include vowels and voiced consonants such as /b/, /d/, /g/, /v/, /z/, and /m/ in English, and possess an F0 that contributes to perceived pitch.

Unvoiced sounds are produced without vocal fold vibration and are characterized by turbulent airflow through specific articulatory configurations. English examples include /p/, /t/, /k/, /s/, /f/, and /h/. These sounds lack regular periodicity and therefore have no discernible F0, but exhibit spectral characteristics that reflect energy distribution across frequencies.

The ability to distinguish between voiced and unvoiced sounds is essential in speech analysis (Greenberg, 1999). This distinction provides valuable insights into speech characteristics, enables speaker identification and language analysis, and is crucial for the development and operation of speech processing technologies.

2.4. Significance in Voiced/Unvoiced Classification

In our research, we utilize both Mel spectrograms and MFCCs for voiced/unvoiced classification in speech signals. These audio representations capture essential spectral characteristics and provide valuable insights into the distinctive attributes of voiced and unvoiced segments. By extracting relevant features from the audio data, we aim to enhance the precision and robustness of the classification task.

In subsequent sections, we introduce the methodology used to train our model on Mel spectrograms and MFCCs extracted from audio signals, and discuss its effectiveness in discriminating between voiced and unvoiced segments.

3. Model Architecture

Our approach employs a Convolutional Neural Network (CNN) model, which has demonstrated exceptional performance in tasks involving image and time-series data, including audio signals. CNNs excel at detecting local patterns or features in data and exhibit translation invariance, making them particularly suitable for our voiced/unvoiced classification task based on time-frequency characteristics of speech signals.

The input audio signal is first processed through two convolutional layers. Each layer employs a Rectified Linear Unit (ReLU) activation function, which introduces non-linearity into the model, enabling it to learn complex patterns. The first layer uses 32 filters, and the second uses 64 filters, both with a kernel size of 3×3. These layers are responsible for extracting relevant features from the input data.

Following the convolutional layers, a max-pooling layer with 2×2 pooling reduces spatial dimensions while retaining the most informative features. This operation reduces computational complexity and helps achieve translation invariance.

To prevent overfitting, dropout layers are incorporated, randomly setting a fraction of input units to zero during training, which improves the model's generalization capability. Through hyperparameter tuning, we determined an optimal dropout rate of 0.4.

The output from the preceding layers is flattened to a single dimension before being input to a fully connected (dense) layer with 128 units and ReLU activation. This layer serves as a classifier for features extracted by the convolutional layers.

The final output layer is a dense layer with a single neuron using sigmoid activation to provide the probability of the input signal being voiced or unvoiced. We use a threshold of 0.5 for binary classification. Threshold tuning on a dev set could improve F1/balanced accuracy; we leave this to future work.

The model is compiled with the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 1×10^{-4}, determined through hyperparameter optimization using Hyperband. We use binary cross-entropy as the loss function, appropriate for our binary classification problem. The model is trained with a batch size of 32 for a maximum of 100 epochs, with early stopping (patience = 5) based on validation loss.

4. Experiments and Datasets

We conduct experiments on two different datasets to validate the effectiveness of our Convolutional Neural Network (CNN) model for voiced/unvoiced classification in speech signals.

Data. The dataset was provided by Saulo Santos; redistribution follows the original corpus licensing terms.

4.1. Dataset 1: MFCC Dataset

The first dataset was created by Saulo Santos, where a baseline model achieved a performance score of 0.82. This dataset contains multiple features extracted from speech signals from three speakers:

- Column 1: F0 reference values in Hz (manually annotated fundamental frequency)
- Column 2 (voiced ground truth): Boolean vector indicating voiced (1) and unvoiced (0) frames based on manual annotation of periodic vocal fold vibration
- Columns 3–16: F0 estimations from several Pitch Detection Algorithms (PDAs)
- Column 17: Harmonics-to-Noise Ratio (HNR) computed using 25 ms analysis windows with autocorrelation
- Column 18: Cepstral Peak Prominence, Smoothed (CPPS) computed following Hillenbrand et al. methodology
- Column 19: Intensity in dB (extracted with Praat using 25 ms analysis windows)
- Column 20: Spectral emphasis following Traunmüller and Eriksson methodology (Traunmüller & Eriksson, 2000)
- Columns 21–40: 20 Mel-frequency cepstral coefficients (MFCCs) computed using the Librosa Python package with 26 Mel filters, excluding C_0, with liftering coefficient 22
- Final column: Group index/file index identifying the original source file

The dataset appears imbalanced (voiced likely dominates continuous speech); we did not apply class weighting in our baseline experiments.

4.2. Dataset 2: Mel Spectrogram Dataset

In addition to the first dataset, we generated a second dataset focused on Mel spectrograms of audio signals. The Mel spectrogram is a powerful feature for speech and audio analysis, capturing both spectral and temporal characteristics. We experimented with different resolutions, generating three versions with varying numbers of Mel bands: 32, 64, and 128.

Like the first dataset, the new datasets include ground truth F0 reference values and voiced/unvoiced labels based on manual annotation. The remaining columns contain Mel spectrogram values for each audio frame, computed using the parameters described in Section 2.1.

4.3. Protocol and Reproducibility

4.3.1. Dataset Splitting and Cross-Validation

Dataset splitting and model training significantly influence predictive performance. We employed 7-fold cross-validation, partitioning the original sample into 7 equal-sized subsamples at the file level to prevent data leakage. One subsample was retained as validation data for testing, while the remaining 6 subsamples served as training data.

Important limitation: With only three speakers in our dataset, the 7-fold cross-validation does not ensure speaker-independent evaluation. Files from the same speaker may appear in both training and validation folds, which limits generalization claims. Future work should employ leave-one-speaker-out (LOSO) evaluation or expand to more speakers for robust cross-speaker validation.

Dataset splitting was implemented using a `split_features_labels` function, distinguishing between feature set (x) and target set (y). Depending on chosen features (`fts`), the feature set varies. For `fts="f0"`, we used features from columns 3–20, whereas for `fts="mfcc"`, we utilized features from columns 21–40.

For data splitting, we employed a KFold strategy using `sklearn.model_selection.KFold` with `n_splits=7`, ensuring balanced splits between training and validation sets. The random seed was fixed to maintain reproducibility.

Split unit. Cross-validation splits are by *file* (never splitting an individual file), not by speaker; thus our results assess within-speaker generalization. Speaker-disjoint evaluation is left for future work.

4.3.2. Sliding Window Processing

To account for the temporal nature of data, we designed a custom Sliding Window Data Generator class capable of creating batches of sliding windows from the dataset as model input. Window size was set to 50 frames (500 ms of audio context), with 50% overlap between consecutive windows.

For window labeling, we use the label of the center frame of each window. Windows spanning file boundaries are discarded to avoid artifacts. Boundary frames between voiced and unvoiced segments are a common error source; future work could use soft labels or exclude ±1–2 frames around transitions. When generating overlapping windows, we ensure that windows from the same file never appear in both training and validation folds to prevent data leakage.

4.3.3. Hyperparameter Optimization

For efficient optimization, we applied Keras Tuner's Hyperband algorithm (Li et al., 2018) to search for optimal model hyperparameters including dropout rates (0.0, 0.2, 0.4, 0.6), learning rate (1×10^{-3}, 1×10^{-4}), and label smoothing (0.0, 0.1). We integrated early stopping callbacks that terminated training if validation loss failed to improve for five consecutive epochs.

Model selection protocol: To avoid overfitting during tuning, a held-out portion of the training fold should be used (e.g., 20%); we treat this as recommended protocol for hyperparameter optimization.

We train our CNN model separately on each dataset and evaluate performance. By comparing model performance across different datasets and configurations, we gain insights into which features and representations are most effective for voiced/unvoiced classification.

5. Results and Discussion

5.1. Model Performance

This section presents a detailed analysis of our deep learning model's performance. We focus on multiple metrics—accuracy, precision, recall, F1-score, and balanced accuracy for both training and validation datasets. These measures provide understanding of the model's capacity to learn from data and its ability to generalize to unseen data.

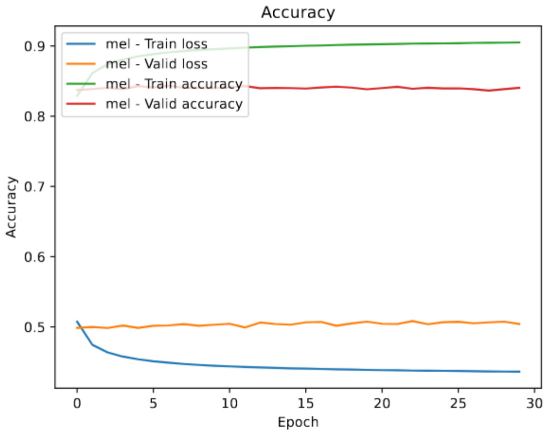

Figure 3. Training and validation loss and accuracy for 32 Mel bands over epochs. Axes: Epoch (x), Accuracy (%, y-left), Loss (cross-entropy, y-right).

Figure 3 illustrates the model's training loss and accuracy over epochs. With 32 Mel bands, the model's training loss converged to 0.42 with training accuracy of 91.1%. The validation loss plateaued at 0.41 with accuracy of 81.3%, indicating some degree of overfitting.

Metric note. Balanced accuracy (BA) is the mean of per-class recalls for the voiced and unvoiced classes, i.e.,

$$\text{BA} = \tfrac{1}{2}(\text{TPR}_{\text{voiced}} + \text{TPR}_{\text{unvoiced}}).$$

Table 1 displays comprehensive model performance metrics on training and validation sets for different numbers of Mel bands. As the number of Mel bands increases, training accuracy improves significantly; however, validation accuracy shows minimal improvement while the training-validation gap increases, indicating progressive overfitting with higher-dimensional representations.

Table 1. Training and validation performance for different Mel band configurations (mean over 7 file-level folds). Higher is better for Acc.

Dataset	Acc_{train} (%)	Acc_{valid} (%)
MEL32	91.1	81.3
MEL64	93.2	81.5
MEL128	96.1	81.8

Table 2 compares classification performance across different feature representations using our CNN model. The MFCC dataset demonstrated superior validation performance compared to Mel spectrogram variants, while Mel spectrogram datasets showed relatively similar performance with marginal improvements as the number of Mel bands increased.

Table 2. Comprehensive performance comparison across feature representations (mean over 7 file-level folds). Higher is better for Acc/BA; lower is better for Loss.

Dataset	Val. Acc. (%)	Val. Loss	Bal. Acc. (%)
MFCC	85.2	0.40	84.8
MEL32	81.3	0.41	80.9
MEL64	81.5	0.42	81.1
MEL128	81.8	0.43	81.4

Across CV folds, MFCCs consistently outperform Mel spectrograms by approximately 3–4 percentage points on validation accuracy and balanced accuracy, suggesting that MFCC features provide more discriminative information for our CNN architecture and dataset size.

5.2. Discussion

Results clearly demonstrate that increasing the number of Mel bands in spectrograms does not lead to significant performance improvement on validation sets. The consistent validation performance across Mel band configurations (81.3–81.8%) suggests that our CNN has reached its learning capacity with the current dataset size. However, larger numbers of Mel bands tend to increase overfitting, as reflected by the growing disparity between training and validation accuracies: 9.8 percentage points for MEL32 (91.1% vs 81.3%), 11.7 pp for MEL64 (93.2% vs 81.5%), and 14.3 pp for MEL128 (96.1% vs 81.8%).

The superior performance of MFCC features can be attributed to their design for speech processing tasks. MFCCs decorrelate log-Mel energies and emphasize spectral envelope characteristics that align well with voicing cues, particularly the separation of source and filter components in speech production. The DCT transformation in MFCCs provides a more compact representation that may be better suited for our limited dataset size.

Mel spectrograms, while preserving local harmonic structure that could theoretically benefit CNNs, may require larger datasets and more sophisticated 2D CNN architectures with greater temporal context to fully exploit their representational capacity.

Typical classification errors likely include unvoiced fricatives (e.g., /s/) being mis-flagged as voiced due to strong high-frequency energy, and breathy or whisper vowels being mis-flagged as unvoiced due to low harmonics-to-noise ratio. The inclusion of HNR and CPPS features in our dataset specifically addresses these challenging cases by providing explicit measures of vocal fold vibration regularity.

6. Limitations

Several limitations should be acknowledged in this study:

- **Small speaker population:** Having only three speakers limits generalization claims and prevents proper speaker-independent evaluation
- **Cross-validation design:** 7-fold CV with 3 speakers allows speaker leakage between folds, compromising generalization assessment
- **Limited metrics:** Primary focus on accuracy; comprehensive evaluation would benefit from precision/recall analysis, confusion matrices, and ROC/PR-AUC
- **Hyperparameter selection:** Potential overfitting to validation performance during model selection
- **Class imbalance:** No explicit handling of the voiced/unvoiced class imbalance
- **Baseline comparisons:** Lack of comparison with traditional methods (energy + ZCR, autocorrelation-based approaches)
- **Window labeling:** Simple center-frame labeling may not optimally handle boundary regions between voiced/unvoiced segments

7. Conclusion

Throughout this study, we have successfully demonstrated the application of convolutional neural networks for voiced/unvoiced classification in speech signals using both Mel spectrograms and MFCC features. Our experiments reveal that MFCC-based features achieve superior validation performance compared to Mel spectrograms across band configurations, likely due to their explicit design for speech analysis tasks and better alignment with voicing discrimination cues.

The consistent validation performance across different Mel band counts (32, 64, 128) suggests that our CNN architecture has reached its learning capacity with the current dataset size, while increasing feature dimensionality primarily leads to overfitting rather than improved generalization. This highlights the importance of matching model complexity to available training data.

Despite the limitations of our experimental setup—particularly the small speaker population and potential speaker leakage in cross-validation—the results provide valuable insights into feature representation choices for voiced/unvoiced classification. The MFCC-based approach showed superior generalization performance, suggesting that cepstral domain representations may be more suitable for this classification task, especially with limited training data.

Looking ahead, future work should explore: (1) expansion to larger, more diverse speaker populations with proper speaker-disjoint evaluation protocols; (2) investigation of more sophisticated temporal models such as Long Short-Term Memory (LSTM) networks (Pradeep et al., 2019) and Transformer-based models like HuBERT (Hsu et al., 2021) or Wav2Vec (Baevski et al., 2020); (3) architectural improvements including batch normalization after convolutions, modest L2 weight decay, global average pooling instead of flattening, and addition of $\Delta/\Delta\Delta$ coefficients for MFCC features; (4) data augmentation strategies such as light time-masking and additive noise for Mel spectrograms; (5) comprehensive evaluation including precision/recall analysis and comparison with traditional baseline methods (energy + zero-crossing rate, autocorrelation-based approaches); and (6) exploration of hybrid approaches combining multiple feature types to leverage the complementary strengths of different representations.

Acknowledgements

I extend my sincere thanks to Marc Evrard for his guidance and mentorship throughout this research project. I also appreciate Albert Rillard's invaluable assistance and insights.

Special thanks to Saulo Santos for his valuable advice on the project and for providing the dataset.

Finally, I am grateful to my fellow researchers for their helpful support throughout this investigation.

References

Ayvaz, U., Gürüler, H., Khan, F., Ahmed, N., Whangbo, T., and Bobomirzaevich, A. A. Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning. *Computers, Materials & Continua*, 71(3):5511–5521, 2022. ISSN 1546-2226. doi: 10.32604/cmc.2022.023278. URL http://www.techscience.com/cmc/v71n3/46499.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

Fant, G. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*. Number 2. Walter de Gruyter, 1971.

Greenberg, S. Speaking in shorthand–a syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29(2-4):159–176, 1999.

Hillenbrand, J., Cleveland, R. A., and Erickson, R. L. Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, 37(4):769–778, August 1994. doi: 10.1044/jshr.3704.769. URL https://doi.org/10.1044/jshr.3704.769.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. Deep learning for monaural speech separation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1562–1566. IEEE, 2014.

Jouvet, D. and Laprie, Y. Performance analysis of several pitch detection algorithms on simulated and real noisy speech data. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1614–1618, 2017. doi: 10.23919/EUSIPCO.2017.8081482.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2015.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: A novel bandit-based approach to hyperparameter optimization, 2018.

Pradeep, R., Reddy, M. K., and Rao, K. S. Lstm-based robust voicing decision applied to dnn-based speech synthesis. *Automatic Control and Computer Sciences*, 53: 328–332, 2019.

Sato, N. and Obuchi, Y. Emotion recognition using mel-frequency cepstral coefficients. *Journal of Natural Language Processing*, 14(4):83–96, 2007. doi: 10.5715/jnlp.14.4_83.

Slaney, M. Auditory toolbox. Technical Report 1998-010, Interval Research Corporation, 1998.

Traunmüller, H. and Eriksson, A. Acoustic effects of variation in vocal effort by men, women, and children. *J. Acoust. Soc. Am.*, 107(6):3438–3451, June 2000.

Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C., and Shamma, S. Linear versus mel frequency cepstral coefficients for speaker recognition. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 559–564, 2011. doi: 10.1109/ASRU.2011.6163888.