Circuits, Systems, and Signal Processing
https://doi.org/10.1007/s00034-024-02794-z

A Quest for Formant-Based Compact Nonuniform Trapezoidal Filter Banks for Speech Processing with VGG16

Cevahir Parlak[1] · Yusuf Altun[2]

Received: 29 August 2023 / Revised: 11 July 2024 / Accepted: 11 July 2024

Abstract
In this text, we discuss the filter banks used for speech analysis and propose a novel filter bank for speech processing applications. Filter banks are building blocks of speech processing applications. Multiple filter strategies have been proposed, including Mel, PLP, Seneff, Lyon, and Gammatone filters. MFCC is a transformed version of Mel filters and is still a state-of-the-art method for speech recognition applications. However, 40 years after their debut, time is running out to launch new structures as novel speech features. The proposed acoustic filter banks (AFB) are innovative alternatives to dethrone Mel filters, PLP filters, and MFCC features. Foundations of AFB filters are based on the formant regions of vowels and consonants. In this study, we pioneer an acoustic filter bank comprising 11 frequency regions and conduct experiments using the VGG16 model on the TIMIT and Speech Command V2 datasets. The outcomes of the study concretely indicate that MFCC, Mel, and PLP filters can effectively be replaced with novel AFB filter bank features.

Keywords Speech processing · MFCC · Mel filters · PLP · Filter banks · Convolutional neural networks

1 Introduction

Filter banks are the driving force in speech processing applications and are constructed over the frequency spectrum of a signal. Triangle, trapezoidal, or Gaussian shapes with different numbers of frequency bands were suggested for the construction of

✉ Cevahir Parlak
cevahir.parlak@fbu.edu.tr

Yusuf Altun
yusufaltun@duzce.edu.tr

[1] Computer Engineering Department, Fenerbahçe University, Istanbul, Türkiye

[2] Computer Engineering Department, Duzce University, Duzce, Türkiye

Published online: 29 July 2024

filter banks in speech processing tasks. Mel filters [111] and their cosine-transform-reduced Mel frequency cepstral coefficients (MFCCs) [4, 16, 27] are still prominent and are the most commonly used features in speech recognition. The Mel scale was introduced by Stevens, Volkmann, and Newman as a perceptual scale of different pitches evaluated as equal from a distance. The human ear is an extremely complicated frequency spectrum analyzer. Although frequency bands are important clues for capturing and understanding speech, the human brain does not rely solely on these features. Human understanding of speech and pattern matching leverages many different aspects, such as back-end ambiance, contextual features, linguistic structures, and language models complemented by the front-end frequency and spatial features, to perform a lightning speed evaluation utilizing the immense parallel processing power of the 100 billion-neuron pattern recognition network of our brain [48, 81].

Human auditory and speech production systems are extremely complicated structures. Myriad theories have been proposed to model the human auditory system. Among them are Mel filter banks, Gammatone filter banks [57], Lyon Auditory filters [73], Ensemble Interval Histograms (EIH) [40], Seneff Auditory filters [107], PLP (Perceptual Linear Prediction) [49], and trapezoidal models [2, 87, 88]. On the other hand, the human vocal tract, which produces sound signals, is another fascinating and intricate mechanism. Vocal folds create a sine wave sound through the air by vibrating at a frequency called the fundamental frequency or pitch; other parts inside the mouth, such as the teeth, tongue, lips, jaw, and even nose, form a filtering mechanism to sculpt the source signal to propagate various phones with different resonant frequencies, which are called formants.

The ERB (Equivalent Rectangular Bandwidth) gammatone model typically extracts 32 critical band filters. The ERB measures the width of the auditory bands throughout the human cochlea and follows a nearly logarithmic scale. Critical bands are also measured in the psychoacoustic experiments of gammatone filter banks. A critical band is a representation of the speech signal at a single auditory nerve unit.

Richard Lyon proposed a cochlear model that describes the human cochlea as a nonlinear filter bank. The cochlear base is stiffer than its apex and is sensitive to high frequencies. The sensitivity decreases from the base to the apex. The Lyon Cochlea model mimics the middle and outer ear in the first stage and behaves like pre-emphasis. In the second stage, an HWR (half-wave rectification) eliminates the negative parts from the input signal, as do the inner hair cells of the ear. In the last part, a cochleagram is formed representing a time–frequency space. Short-time autocorrelation (STA) is applied to the outputs to create a cochleagram for the nonstationary speech frames. A correlogram is a 3D time–frequency-lag space and involves cochleagram representation. For 16 kHz sound signals, Lyon's cochlear model facilitates 86 filters.

Seneff's model comprises 40 filters representing the motion of the basilar membrane and auditory nerve response. Synchrony outputs and mean rates are two outputs of the Seneff cochlear model design. The mean rates are derived from the envelope of the stage output and can be considered the spectral magnitude representations. On the other hand, synchrony outputs aim to discover the center frequencies of consonants (stops, fricatives, and sonorants).

The EIH (Ensemble Interval Histogram) is another hearing model suggested by Ghitza in 1992. The EIH model is quite similar to the Seneff model in the beginning stage; however, it constructs 85 filters compared to the 40 filters used in the Seneff model. The second part of the EIH generates histograms per channel. The final part aggregates the histograms of the second stage, known as the Ensemble Interval Histogram.

The PLP (Perceptual Linear Prediction) model was proposed by Hermansky and comprises 24 filters based on the Bark scale proposed by Zwicker as a psychoacoustic hearing model. Perceptual linear predictive coding leverages the cubic-root intensity-loudness power law and equal loudness curves to flatten the spectral magnitudes of the critical bands. PLP also uses an all-pole autoregression model to simulate the human vocal tract to provide a clear representation of the auditory spectrum. This allows the PLP to simulate human hearing better than the LPC [12]. PLP is less sensitive to noise and computationally more efficient than linear predictive coding. RASTA-PLP (**R**el**A**tive **S**pec**T**r**A**l PLP) [50] was launched to enhance the efficiency of PLP for communication transmission channels.

The Mel filter bank is the most common filter bank used in speech-processing applications. Mel filters have 40 triangle-shaped frequency bands that crossover with one another. However, there is no concrete agreement among researchers over the frequency regions of these bands. The frequency regions of the triangular bands of the Mel filters may be decided according to the applications (music, emotion, speech, speaker, gender, etc.). These triangular-shaped filter banks try to mimic the human auditory system. Their cousin Mel-Frequency Cepstral Coefficients (MFCCs) are computed by applying the DCT (Discrete Cosine Transform) to the logarithmic magnitude frequency spectrum of 40 Mel filters. Davis, Mermelstein, and other researchers claim that MFCC can be considered an application of principal component analysis (PCA) to the logarithmic power spectrum. While MFCC is very successful in speech recognition implementations, it is ineligible for speech synthesis due to the impossibility of reversing the DCT operation. It is also highly susceptible to noise.

Studies of the imitation of the human vocal tract date back to the late eighteenth century by Christian Kratzenstein, who explained the differences between /a/, /e/, /i/, /o/, and /u/. The quest continued with Wolfgang von Kempelen, Charles Wheatstone, Alexander Graham Bell, and Herman von Helmholtz [36, 104]. The first speech synthesis devices were introduced by Homer Dudley and Walter Lawrence [34, 68]. Gunner Fant introduced the first cascade formant synthesizer, followed by Allen, Umeda, Holmes, Rosen, and Klatt with MITalk [7, 35, 54, 65, 99, 118]. Vowel and consonant formulations have also been extensively studied by many researchers, including Peterson, Barney, Wells, Lieberman, Ladefoged, Johnson, Rabiner, Hillenbrand, Assman, Klautau, Coleman, Kewley, Cox, Bernard, Hagiwara, Harding, Picone, Stevens, Huckvale, Kidd and Jurafsky [11, 14, 22, 24, 43, 45, 51–53, 55, 58–61, 63, 66, 67, 72, 90, 92, 94, 95, 112, 121]. Linguistics and phonetics have also contributed to understanding the resonant frequencies of speech signals via vocal tract articulatory movements for speech synthesis and analysis [6, 32, 41, 47, 84, 126]. Vocal tract resonant frequencies of speech signals are called formants and are formed during the passage of air through the vocal path [5, 29, 56, 79, 89, 113]. VOT (Voice Onset Time) is a significant factor in the identification of stop consonants [19, 20]. Nasal consonants have their own special structures [46]. The formation of the /r/ sound is special, and its formants are strongly influenced by accompanying vowels [123]. The application areas of speech processing are too diverse, including gender, age group, and speaker recognition [78, 102]. Speech phones, particularly vowels, have also been investigated in the country domain, and there are several studies on Turkish vowels [10, 15, 17, 64, 70].

The acoustic simulation of the vocal tract can be implemented as an approximation by lossless two-tube or three-tube models [8, 9, 18, 26, 30, 35, 74, 77, 96, 97, 101]. The human ear is more discriminative but less sensitive at lower frequencies, whereas at high frequencies, it is less discriminative but more sensitive. A 3000 Hertz sound can be perceived better than a 100 Hertz sound with the same amplitude. However, in the low-frequency region, it is easier to discern different frequencies. This phenomenon constitutes the foundation of Equal Loudness Curves [37, 98, 114]. The human frequency spectrum is linearly spaced below 1200 Hz and logarithmic beyond that region. Our ear is a logarithmic frequency analyzer, and its working range is amazing. It has a 3.6 Hz frequency resolution between the 1000–2000 Hz band under ideal test conditions [80, 86, 106, 122, 125].

Given this background, we need to explain the necessity of proposing novel filters. Novel AFB filters are designed to compete with Mel filters, PLP filters, and MFCC features, which are the most widely used representations of speech processing applications. Mel filters contain 40 bands, and MFCC uses 13 of these 40 bands via discrete cosine transform. The problem with Mel filters is that they use too many filters, thereby indicating overfitting and computational and temporal overloads despite the high accuracy rates. MFCC has smaller coefficients; however, the performance of MFCC, particularly in deep learning applications, is unsatisfactory. In our study, the PLP is implemented with 21 subbands. The proposed AFB filters will provide a mechanism in between which a smaller number of coefficients will be used than for the Mel filters and will provide accuracies comparable to those of the Mel and PLP filters. Currently, AFB features are designed to include only 11 trapezoidal frequency bands. In Sect. 3, we comprehensively delineate the proposed AFB filters.

The remainder of this manuscript is organized as follows: Sect. 2 reviews related studies, Sect. 3 provides a detailed explanation of the proposed AFB filter banks, Sect. 4 discusses the speech datasets used in our experiments, Sect. 5 addresses the convolutional neural network used in the experiments, Sect. 6 presents the experimental results, and Sect. 7 finalizes the paper with the conclusions and future directions.

2 Related Works

In this text, we run experiments on the SCD (Speech Command Dataset) [120] and TIMIT [39] datasets. There are numerous studies on these widely used datasets, including that of Andrade et al. [28], who studied a convolutional recurrent neural network with attention on SCD v1 and v2. They achieved 93.9% accuracy on v2 for the 35-command recognition task with an attention-RNN model. They extracted 80-band Mel-scale features with 1024-point Fast Fourier Transform (FFT) frames and 128-point overlapping windows. The model applies a set of convolutions to the feature vector followed by a set of 2 bidirectional LSTM nodes. The LSTM output is then passed through 3 dense layers.

In [115], Toth proposed maxout neurons in convolutional neural networks as an alternative to the rectified linear unit function. They conducted experiments on the TIMIT dataset and achieved outstanding phone error rate performances. Experiments were run with 40-dimensional Mel filter bank features plus the energy of the frame. Delta and double delta coefficients are also computed to yield 123 features in total. This paper outperformed previous works by revealing a 16.5% phone error rate using the hierarchical CNN model. The author also tested the Hungarian Broadcast News Corpus as a large vocabulary continuous speech recognition task. The Szeged dataset contains 28 h of speech data from Hungarian TV channels. In the experiments, the training set utilized 22 h of data, 2 h of data were used as the validation set, and 4 h were allocated for testing purposes. In this second experiment, the proposed CNN model achieved the best performance, with a 15.5% phone error rate.

In [33], Dridi and Ouni proposed CGDNN (convolutional gated deep neural network) and conducted phoneme recognition experiments on the TIMIT dataset. The result is a 15.72% phone error rate using 40-dimensional Mel filter bank features with their delta and double delta derivatives.

In [71], Li and Zhou tested a single-layer softmax, a 3-layer fully connected DNN, and a convolutional neural network on SCD v1 for KWS (keyword spotting task). The speech wave files are processed with 30 ms framing and a stride of 10 ms to extract a 40-dimensional feature set. They used only 6 words from the SCD v1. The CNN model demonstrated extraordinarily high performance over the softmax DNN and vanilla RNN models, with accuracies of 94.5%, 71.9%, and 56.7%, respectively.

Berg et al. introduced the keyword transformer [13] to SCD v1 and v2. They used 40 Mel filters with an 80:10:10 train:validation:test set split along with data augmentation and preprocessing. They achieved 97.27% by the multihead attention-RNN model, 97.53% by KWT-2, and 97.74% by the KWT-2 distillation model on the SCD v2 with all 35 keywords.

Trinh et al. [116] experimented on SCD v2 and proposed a novel augmentation method called ImportantAug, which adds noise to the unimportant parts of speech data. They used additional noise with importance maps. They achieved a 6.7% error rate without augmentation, 6.52% with conventional noise augmentation, and 5.00% with the proposed ImportantAug method.

3 Proposed AFB Filter Banks

The structure of human hearing has been studied extensively, and many models have been proposed to imitate the auditory system. The Mel filter bank and MFCC have been used for over half a century for speech processing, and time is ripe to replace them with better features to represent speech signals. In this study, a novel filter bank strategy named acoustical filter banks (AFB) is proposed for speech processing applications to replace Mel filters, PLP filters, and MFCCs. The foundations of novel AFB filters rely heavily on the formant regions of vowels and consonants. The novel features contain only 11 marginally overlapping trapezoidal frequency subbands, as delineated in Table 1 and graphed in Fig. 1. They are less expensive to compute, provide a more compact representation of the data to obtain more information about

Table 1 Frequency bands of the novel AFB filters

Filter no.	$f1$	$f2$	$f3$	$f4$
1	100	130	230	270
2	230	270	370	400
3	350	400	500	550
4	500	550	750	800
5	750	800	1000	1100
6	1000	1100	1300	1400
7	1300	1400	1700	1800
8	1700	1800	2100	2300
9	2100	2300	2700	3000
10	2700	3000	3700	4000
11	3700	4200	5500	6000

The numbers are in Hertz

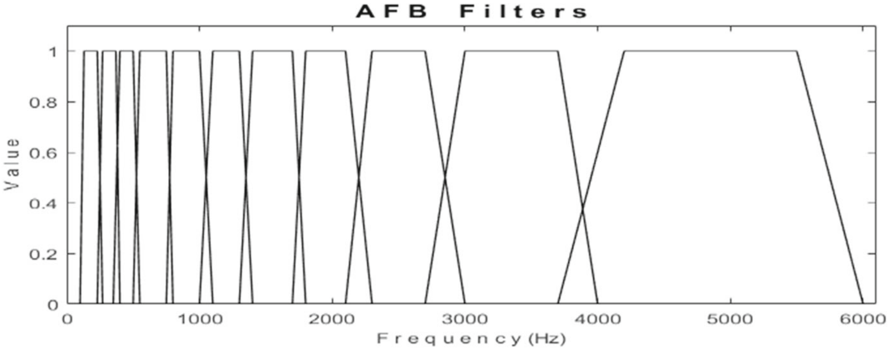


Fig. 1 Picture of the proposed Acoustic Filter Banks

the underlying dynamics, and offer enhanced interpretability compared to Mel filters or MFCC. In Fig. 2, we sketch the graphs of the AFB, MFCC, and Mel filters of the vowels /i/, /u/, and /a/ respectively. In speech processing, we divide the speech signal into windowed frames of a certain length, such as 25-ms frames with 10-milisecond overlaps. Therefore, processing a single phone requires several frames. In Fig. 2, each line in the graph of a vowel represents these windowed frames of the vowel signal. Although frequency information, namely, the first and second formants ($\mathcal{F}1$, $\mathcal{F}2$), can strongly represent vowel sounds, consonants cannot be segregated solely by employing frequency regions. The formation type of the phone is very effective in determining the final consonant phone. It is possible to create very different consonants with the exact same spectral structure. In Fig. 3, the time-domain signals and spectrograms of the words "hissing" and "hitting" are shown. Here, we make a trick on the time-domain signal of "hissing" and silence the yellow-marked portion without interfering with

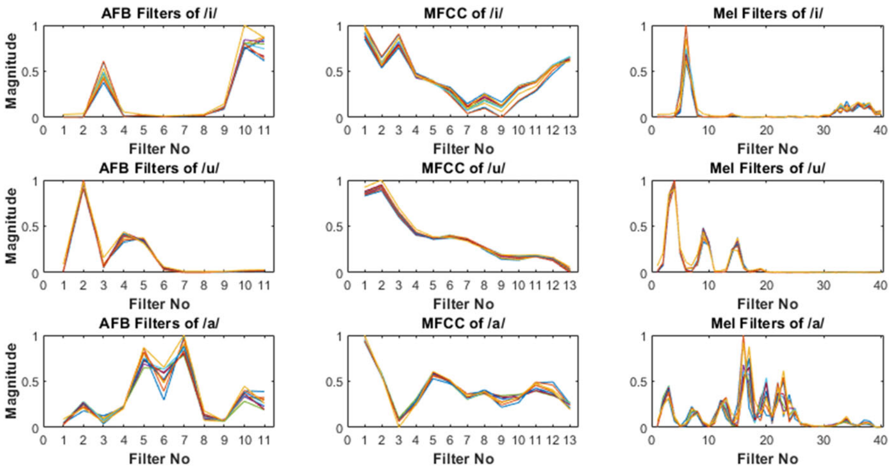
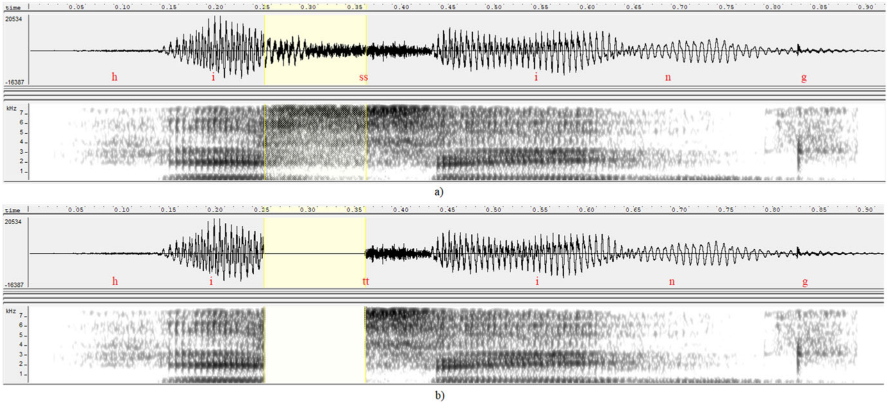


Fig. 2 Visualization of the power spectra of the AFB, MFCC, and Mel filters for the vowels /i/, /u/, and /a/. Each line represents a windowed frame of the vowel



a)

b)

Fig. 3 The wave signal and frequency spectrogram of the words **a** “hissing” and **b** “hitting” after silencing the yellow marked region of the word “hissing”. Here, we do not modify the frequency domain. We silence only the region marked in yellow (Color figure online)

the frequency domain at all. As shown in Fig. 3, the yellow marked region is a part of the consonant /s/ in the word “hissing”. This will create a silent part immediately before the second half of the phone /s/, which is commonly called VOT [19, 20], and the whole signal will be heard as “hitting” instead of “hissing”. This is one of the difficulties of speech recognition that makes it incredibly formidable along with the different lengths of each phone, from person to person or even within the same person, and varying phone boundary regions. Delta acceleration coefficients, Haar wavelets [42], or other change point detection methods can be helpful for identifying such sharp changes across frequency bands.

The construction of AFB filters heavily depends on the formant regions of vowels and consonants. For this purpose, we used the formant regions of vowels from current studies. Although vowel theory is well established, consonants need more attention and are difficult to observe since they are posing extreme complexity due to their diverse articulatory formations. In contrast to the studies about vowels, there are considerable differences among researchers about consonant bands, and consonant band studies are scarce compared to vowel studies [19, 20, 25, 38, 43, 46, 65, 66, 78, 89, 102, 121, 123, 124]. We implemented millions of binary classification experiments to determine the most discriminative frequency bands between the acoustical neighbors and similar phones by employing all possible frequency subband pairs. These experiments helped us to explore the different frequency regions between vowels and consonants, leading to the construction of fine-tuned subbands of AFB filter banks.

At the outset of our study, we set the upper frequency boundary of the AFB's 11th filter to 5000 Hz. However, upon a more detailed investigation, we found that the phone /s/ (and arguably the phones /z/, /ʤ/, /tʃ/, /ʃ/, /ʒ/, /k/, /g/, /t/, /d/) has wider spectral bandwidths spanning from 3000 Hz up to 7000 Hz, particularly when accompanied by the vowel /iy/ or /ih/. We experimented with 5500 Hz, 6000 Hz, and 7000 Hz boundaries, and the results are nearly identical for 6 kHz and 7 kHz, while 5500 Hz slightly lags behind. Therefore, for the sake of keeping the spectrum as narrow as possible, we selected 6000 Hz as the ending boundary of the AFB filters. We did not add a new frequency band here because we did not observe any distinct frequency region between any other phone. Another interesting finding is that the arrangement of the input features is highly effective in terms of performance. The speech signals are inherently 1D, and when they are fed into a 2D-CNN, they should be converted to a 2D matrix. The performance becomes best when they are in the matrix form of $(frame\_count \times feature\_count)$ instead of selecting arbitrary rows and columns.

From Fig. 3, we can observe that AFB filters unearth the nature and structure of these sounds better than do the MFCC and Mel filters. Mel filters can be interpreted as better than MFCCs; however, they fall short of AFB filters. It is quite difficult for the MFCC to find any evidence about the structure of phones. AFB filters can be regarded as a compact view of Mel filters, emphasizing distinct passband regions for phonetic discrimination. The phones that are acoustic neighbors should have a disparate (passband) region where the phone is perceived exactly as it is. There are also crossover overlapping (transition band) regions between the acoustic neighbors where the phone can be perceived as either of them. Humans usually mix the pronunciation of acoustic neighbors or similar phones such as /u/-/o/, /o/-/a/, /u/-/ɯ/, /e/-/i/, /s/-/z/, /ʤ/-/tʃ/, /ʃ/-/ʒ/, /k/-/g/, /p/-/b/, and /t/-/d/. This idea is supported by numerous studies with the highest error rates for similar phones in confusion matrices [51, 90]. Therefore, it is difficult to achieve perfect discrimination between phones. Instead, such burdens should be handled by language models, which can determine the closest matching words or sentences.

AFB filters can be represented using a nonuniform filter bank summation of the short-time Fourier transform as follows:

$$\omega_k = \frac{2\pi}{N}, k = 0,\ 1,\ 2, \ldots, N - 1$$

$$\omega_k = \omega(n)$$

$$\mathrm{H}_k\left(e^{j\omega}\right) = W_k\left(e^{j(\omega-\omega_k)}\right)$$

$$h_k(t) = a^{-k/2}h\left(b^{-k}t,\right)$$

$$\mathrm{H}_k\left(e^{j\omega}\right) = a^{k/2}H\left(e^{j\omega b^k}\right)$$

The nonuniform bandwidths of subbands in AFB and nonuniform decimation are typical components of wavelet filters, where all frequency responses are obtained via frequency scaling instead of frequency shifting via short-time Fourier transform. Note that nonuniform subbands are highly compatible with the human hearing system. As an alternative for nonuniform filter bank summation, we can use *Fejér–Korovkin* [82] wavelet filters to construct marginally overlapping trapezoidal frequency domain filters. Fejér–Korovkin kernel (K_m) is defined by:

$$K_m(\xi) = \left\{ \begin{array}{l} \frac{2sin^2(\pi/(m+2))}{m+2}\left[\frac{cos((m+2)x/2)}{cos(\pi/m+2))-cos(\xi)}\right]^2, x \notin \mp\frac{\pi}{m+2} + 2\mathbb{Z}\pi \\ (m+2)/2, x \in \mp\frac{\pi}{m+2} + 2\mathbb{Z}\pi \end{array} \right\}$$

K_m can be written in the form of

$$K_m(\xi) = 1 + 2\sum_{k=1}^{m}\theta_m(k)coskx$$

where

$$\theta_m(k) = \frac{\left[(m-k+3)sin\frac{k+1}{m+2}\pi - (m-k+1)sin\frac{k-1}{m+2}\pi\right]}{2(m+2)sin\left(\frac{\pi}{(m+2)}\right)}$$

The *Fejér–Korovkin* filter is expressed as follows:

$$\left|h_0^m(\xi)\right|^2 = \frac{1}{2\pi}\int_{-\pi/2}^{+\pi/2} K_m(\xi-u)du$$

Nonuniform m-channel quadrature mirror filters or cosine modulations ensure perfect reconstruction of signals when constrained to a paraunitary polyphase matrix with significant simplification even in multirate systems. Cosine-modulated analysis and synthesis filters can also represent AFB filters [119]. A comprehensive discussion of these filters is beyond the scope of this manuscript. In Fig. 4, *Fejér–Korovkin* filters and their normalized frequency magnitude responses are depicted. As seen from c)

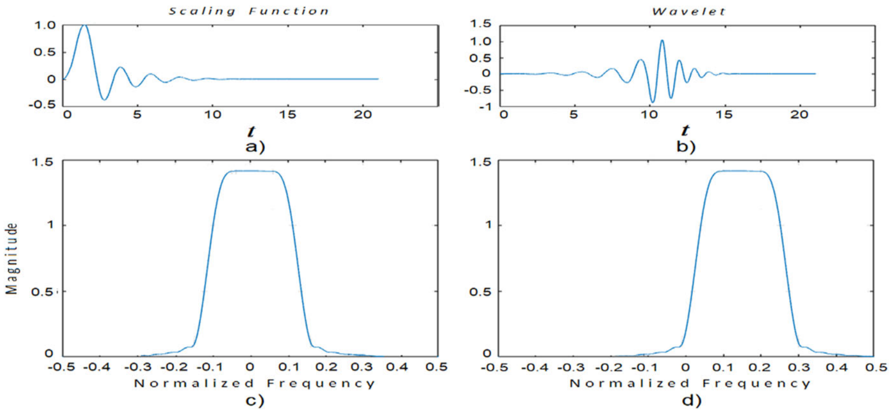


Fig. 4 Fejér–Korovkin filters and their frequency magnitude responses

and d) of Fig. 4, they produce a near-trapezoidal-like frequency domain bandpass filter bank. AFB filters can be used as analysis filters as well as synthesis filters for the perfect reconstruction of the signal.

Vowels and consonants can also be studied and simulated using acoustic tube models. An approximate 3D drawing of the vocal tract simulator with 4 concatenated tubes and the nasal cavity is shown in Fig. 5. In a closed acoustic tube, sound waves are

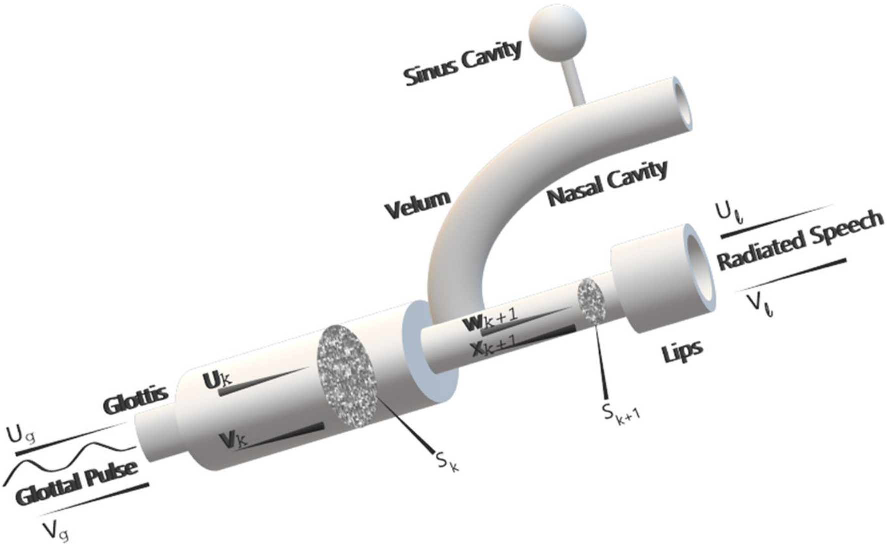


Fig. 5 Simulation of the vocal tract using concatenated acoustic tubes

governed by the following equations:

$$-\frac{\partial \boldsymbol{p}}{\partial \boldsymbol{x}} = \frac{\boldsymbol{\rho}}{\boldsymbol{S}}\frac{\partial(\boldsymbol{u})}{\partial \boldsymbol{t}}$$

$$-\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} = \frac{\boldsymbol{S}}{\boldsymbol{\rho c}^2}\frac{\partial(\boldsymbol{p})}{\partial \boldsymbol{t}}$$

In this equation, the cross-sectional surface area (S_k, S_{k+1}) of the k^{th} tube is assumed fixed such that $S(x, t) = S$, p is the pressure, u is the velocity of the wave at the x position and t time, c and ρ are the density of air and the velocity of the sound wave in the acoustic tube, respectively. Nonuniform lossless tubes have no closed-form solutions. Solving the wave equations remains difficult even when the above assumptions are used. The solution of the above equation for the k^{th} tube can be written as follows:

$$p_k(x, t) = \frac{\rho c}{S_k}\left[u_k^+\left(t - \frac{x}{c}\right) + u_k^-\left(t + \frac{x}{c}\right)\right], 0 \le x \le \ell_k$$

$$u_k(x, t) = u_k^+\left(t - \frac{x}{c}\right) - u_k^-\left(t + \frac{x}{c}\right), 0 \le x \le \ell_k$$

where $u_k^+(t - x/\text{c})$ is the forward wave, $u_k^-(t - x/\text{c})$ is the backward wave, and ℓ_k denotes the length of the k^{th} acoustic tube. Using the flow and pressure continuity at the junction of the tubes, we can derive the following equations in matrix form:

$$flow\ continuity(U_k - V_k) = (W_{k+1} - X_{k+1})$$

$$pressure\ continuity\frac{\rho c}{S_k}(U_k + V_k) = \frac{\rho c}{S_{k+1}}(W_{k+1} + X_{k+1})$$

$$\begin{bmatrix} 1 & -1 \\ S_{k+1} & S_{k+1} \end{bmatrix}\begin{bmatrix} U_k \\ V_k \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ S_k & S_k \end{bmatrix}\begin{bmatrix} W_{k+1} \\ X_{k+1} \end{bmatrix}$$

$$\begin{bmatrix} U_k \\ V_k \end{bmatrix} = \frac{1}{2S_{k+1}}\begin{bmatrix} S_{k+1} & 1 \\ -S_{k+1} & 1 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ S_k & S_k \end{bmatrix}\begin{bmatrix} W_{k+1} \\ X_{k+1} \end{bmatrix}$$

By defining r_k as the amount of $u_{k+1}^-(t)$ that is reflected at the junction point,

$$r_k = \frac{S_{k+1} - S_k}{S_{k+1} + S_k}, (-1 \le r_k \le +1)$$

$$\begin{bmatrix} U_k \\ V_k \end{bmatrix} = \frac{1}{1 + r_k}\begin{bmatrix} 1 & -r_k \\ -r_k & 1 \end{bmatrix}\begin{bmatrix} W_{k+1} \\ X_{k+1} \end{bmatrix}$$

We can convert this time delay to multiplication using the z-transform with $z^{-1/2}$:

$$U_k(z) = z^{-1/2}X_{k+1}(z)$$

$$V_k(z) = z^{+1/2} W_{k+1}(z)$$

For an n-segment concatenated tube with the assumptions of $V_\ell = 0$ (no reflection at the tip of the mouth), V_g can be ignored due to absorption by the lungs.

$$\begin{bmatrix} U_g \\ V_g \end{bmatrix} = z^{1/2} \begin{bmatrix} 1 & 0 \\ 0 & z^{-1} \end{bmatrix} \begin{bmatrix} W_{k+1} \\ X_{k+1} \end{bmatrix} = \frac{z^{n/2}}{\prod_{k=0}^{n}(1+r_k)} \prod_{k=0}^{n-1} \begin{bmatrix} 1 & -r_k z^{-1} \\ -r_k & z^{-1} \end{bmatrix} \times \begin{bmatrix} 1 \\ -r_n \end{bmatrix} U_\ell$$

Hence, the transfer function is defined as:

$$V(z) = \frac{U_l}{U_g} = \frac{G z^{-n/2}}{1 - \alpha_1 z^{-1} - \alpha_2 z^{-2} - .... - \alpha_p z^{-n}}$$

where G is the gain, $z^{-n/2}$ is the delay across the vocal tract, and the denominator part of the transfer function equation is an all-pole filter with the order of n. With the help of closed acoustic tubes, we can explore the formants of vowels and consonants.

There are many studies on the formants of vowels and consonants. Some studies, such as the Hillenbrand and North Texas vowel datasets, provide f_0, $\mathcal{F}1$, $\mathcal{F}2$, $\mathcal{F}3$, and even $\mathcal{F}4$ values. Currently, most of them agree on the frequency regions of vowels, except for some rare cases such as /uu/ and /al/, as depicted in Figs. 6 and 7 of the Hillenbrand and North Texas vowel datasets. The first formant together with the second

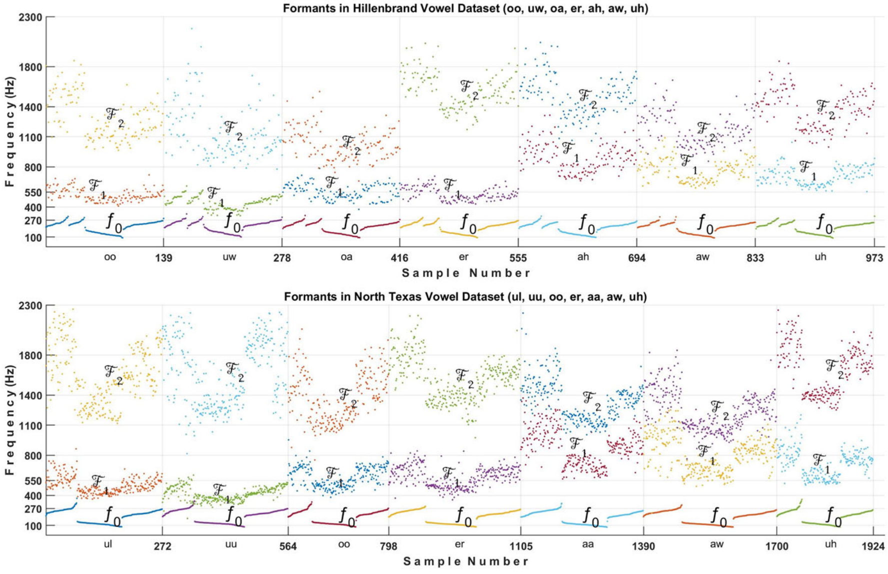


Fig. 6 Comparison of fundamental and formant frequency values of the Hillenbrand and North Texas datasets for the vowels /oo/, /ow/, /oa/, /er/, /ah/, /aw/, and /uh/. The boy, girl, woman, and kid classes are sorted by the f_0 value in ascending order, and the man class is sorted by the f_0 value in descending order for better visual discrimination

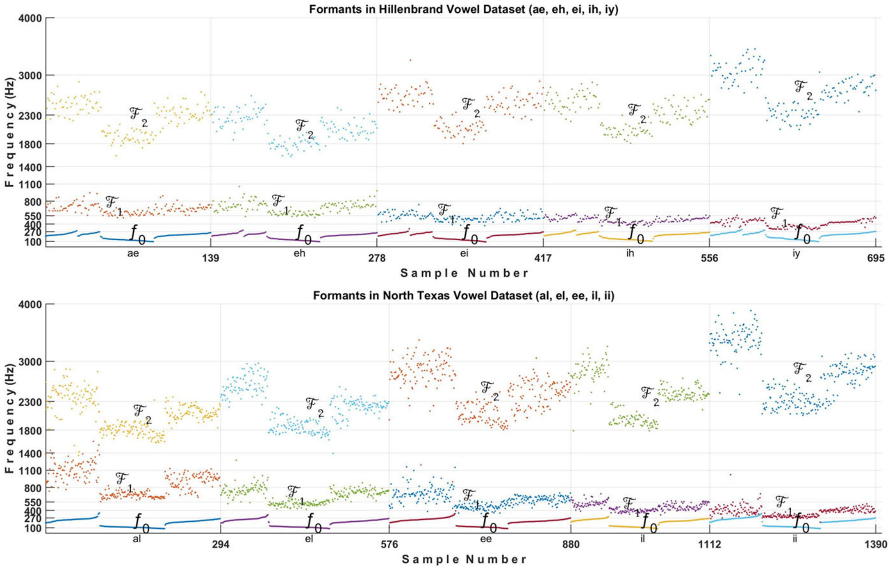


Fig. 7 Comparison of fundamental and formant frequency values of the Hillenbrand and North Texas datasets for the vowels /al/, /el/, /ee/, /il/, and /ii/. The boy, girl, woman, and kid classes are sorted by the f_0 value in ascending order, and the man class is sorted by the f_0 value in descending order for better visual discrimination

formant can adequately represent the vowels. I would suggest that in some vowels such as /iy/, we may not need two formants. In phone /iy/, $\mathcal{F}2$ can sufficiently represent the vowel. Philips et al. also studied single formants [91].

It is well known that when a speech signal is high-pass-filtered above 500 Hz, the intelligibility is nearly intact except for the level of loudness due to the loss of fundamental frequency components, which is almost always below 500 Hz [31]. Interestingly, 500 Hz is the first resonant frequency of a neutral vocal tract shape that produces the vowel /e/. I would make a suggestion that the formants that are below 500 Hz should be discarded as they are not required for the perception and intelligibility of the phones. This region below 500 Hz is not significant for speech recognition. For instance, when a high-pass filter with a cutoff frequency of 1000 Hz is applied to the sentence "She sees seas", it is still intelligible except for some levels of loudness loss. All the phones in this sentence have formant bands above 1000 Hz except for the first formant of /iy/. Even though a high-pass filter is applied twice to remove possible artifact remnants, it does not lose its intelligibility, which cannot be explained by means of missing $\mathcal{F}1$, such as the long-disputed missing f_0 concept. However, we should keep in mind that removing a frequency component below 500 Hz does not necessarily mean the removal of its perception by the human brain, as in the case of long-standing missing fundamental dilemma [85, 103, 105, 117]. Another interesting phenomenon related to speech is the difference between genders. The voices of men, women, boys, and girls have different characteristics [23, 83]. Speech processing techniques also use a variety of features, including wavelet and wavelet packet transforms [75, 108].

The Hillenbrand and North Texas datasets provide valuable information about the formants of vowels; hence, we opted to present them in Figs. 6 and 7. Each vowel class is depicted in the order of boy-girl-man-woman in the Hillenbrand dataset and kid-man-woman in the North Texas dataset in Figs. 6 and 7. In the figures, the boy, girl, woman, and kid classes are sorted by the f_0 value in ascending order; however, the man class is sorted by the f_0 value in descending order for better visual discrimination. Kids are at ages 3, 5, and 7 in the North Texas dataset, whereas there is no age information in the Hillenbrand dataset. Hillenbrand dataset contains 1668 vowel samples (540 males, 576 female, 324 boys, 228 girls) and North Texas datasets contains 3314 vowel samples (972 kids, 1232 males, 1110 females). In Figs. 6 and 7, the vertical axis denotes the frequency edges of the AFB filters, and the horizontal axis denotes the vowel ARPABET class with the sample number.

In North Texas and Hillenbrand, $\mathcal{F}1$ and $\mathcal{F}2$ of /aa/ and /aw/ phones are nearly in the same regions. In the North Texas dataset, $\mathcal{F}2$ is higher than that in the Hillenbrand dataset for /uh/, /ul/, and /uu/ phones. In North Texas, $\mathcal{F}2$ of /ul/ and /uu/ is greater for children and women, which may be due to mispronunciation and mislabeling, particularly for kids, accents, formant calculation errors, or outliers. $\mathcal{F}1$ is in the same place in all (/oo/, /uw/, /oa/, /er/, /ah/, /aw/, /uh/). The phone /oo/ is sometimes pronounced like /u:/, as in the case of hue with the tongue slightly forward, thus raising $\mathcal{F}2$. In the North Texas dataset, the $\mathcal{F}2$ of /oo/ is greater than the $\mathcal{F}2$ of /oa/ of the Hillenbrand vowel dataset. Examining the /oa/ and /er/ phones in the Hillenbrand vowel dataset and /oo/ and /er/ in the North Texas vowel dataset, we observe that the $\mathcal{F}2$ formant of /er/ is shifted one level upward compared to the /oa/ of Hillenbrand or /oo/ of the North Texas dataset, while $\mathcal{F}1$ remains nearly on the same band. This is a great clue for exploring articulatory movements. There are two main vocal tract articulations, namely, tongue and lip movements. Jaw movement corresponds to movement of the tongue up or down. With regard to the pronunciation of /er/, we moved our tongue slightly further than we did with respect to /oa/ and /oo/. Therefore, we can conclude that the tongue forward raises the F2 formant. Lip forwarding is the other articulatory movement and lowers the $\mathcal{F}1$ formant. In Fig. 7, $\mathcal{F}1$ is in the same place in all (/al/, /el/, /ee/, /il/, /ii/) vowels except /al/ of kids and women. $\mathcal{F}2$ is approximately in the same place in all (/al/, /el/, /ee/, /il/, /ii/). The fundamental frequency f_0 is the same for all vowels, with nearly perfect agreement in both datasets. The formant structures of /ih/, /iy/, /ii/, /ae/, /eh/, /ei/, /al/, /el/, and /ee/ are very clearly identified in both the Hillenbrand and North Texas datasets. Almost all studies on speech analysis agree on this issue with subtle differences [5, 7, 10, 11, 14, 15, 17, 24, 35, 51, 52, 56, 60, 61, 64, 65, 70, 72, 79, 89, 90, 113, 121]. From Figs. 6 and 7, we can observe how elegantly the formants align with the frequency regions of AFB filter banks. The diphthongs /oy/, /ay/, and /ey/ should be considered as the concatenation of /oa/, /aa/, and /eh/ with /iy/, respectively. Therefore, the frequency spectra of /oy/, /ay/, and /ey/ are strongly affected by the frequency spectrum of /iy/ due to the time-blindness of the FFT. In TIMIT, /ao/ pronunciation is sometimes like /ow/ and sometimes like /aa/. The formant scatter plots do not provide enough information for vowel discrimination; hence, we tested other plot types. The histograms and boxplots of the AFB filters of vowels in the Hillenbrand, North Texas, and TIMIT datasets are presented. Boxplot representation produces better visualization for the discrimination of vowels regarding the median points.

As we stated, consonants cannot be separated solely by means of formants. Some consonants have the same formant structure, but their articulatory formations are different. We can convert a consonant to another consonant by applying appropriate low-pass, high-pass, bandpass, or bandstop filters. The phone /ʒ/ can be converted to /s/ by applying a high-pass filter above a 3000 Hz cutoff frequency when accompanied by /u/; on the other hand, when flanked by phone /i/, the cutoff frequency may increase up to 4000 Hz. Removing the frequency bands between 500 and 3000 Hz using a bandstop filter will convert /ʒ/ into /z/. When accompanied by /i/, this bandstop region may extend between 500 and 4000 Hz. It is also possible to convert /ʒ/ into /ʃ/ by applying a bandstop filter between 0 and 2000 Hz. However, if /ʒ/ is coupled with /i/, this bandstop region will extend between 0 and 3000 Hz. There are some other conversions by means of adding or removing VOT before some phones. As illustrated in Fig. 3, we can silence the first half of the /s/ and convert it to /t/. Conversely, deleting this VOT before /t/ will convert /t/ into /s/. The same changes apply to the /ʃ/-/tʃ/ and /ʒ/-/ʤ/ pairs. The phone /ʃ/ can be converted to /s/ by removing the frequency region between 1500 and 3000 Hz. We can apply similar transformations for the /k/-/g/, /p/-/b/, /t/-/d/ and /m/, /n/, /l/, /r/, /f/, /v/ consonants. The vocal tract is also accompanied by the nasal cavity as a parallel sound wave transmission line. The nasal cavity affects the speech signal by adding a zero or anti-formant over the 1000 Hz frequency region. Therefore, nasal phones (/m/, /n/) have very little high-frequency energy. The large surface of the nasal cavity causes greater thermal loss and viscous friction, leading to larger bandwidths for nasal resonances.

In our study, we relied on the formant regions of vowels and consonants to construct AFB filter banks; however, a detailed comprehensive discussion of phones exclusively of consonants is completely beyond the scope of this paper. Interested readers may find further information in the related references of this manuscript.

In Figs. 8, 9, and 10, we present the histograms of Hillenbrand, North Texas, 20 TIMIT vowels, and 24 TIMIT consonants. We chose to represent all vowels in the TIMIT instead of the mapped ones to clarify the differences between them, if any. As seen from the histograms of Fig. 9 for the TIMIT dataset, there are significant differences between the mapped /ah/-/ax/-/axh/, /uw/-/ux/, /ao/-/aa/, and /er/-/axr/ whereas the difference between /ih/-/ix/ is not noticeable.

Formant plots and magnitude histograms do not provide sufficient information for the discrimination of vowels; therefore, we decided to construct boxplots. Boxplot representation provides a better understanding when the median value is taken as the pivot point. Boxplots of the Hillenbrand and North Texas datasets are depicted in Fig. 11, boxplots of the TIMIT vowels are shown in Fig. 12, and TIMIT consonant boxplots are shown in Fig. 13.

Another advantage of boxplots is that they can be easily and more correctly computed from filter bank magnitudes than can formant calculations. Although we present the histograms and boxplots of consonants in the TIMIT dataset, we should emphasize

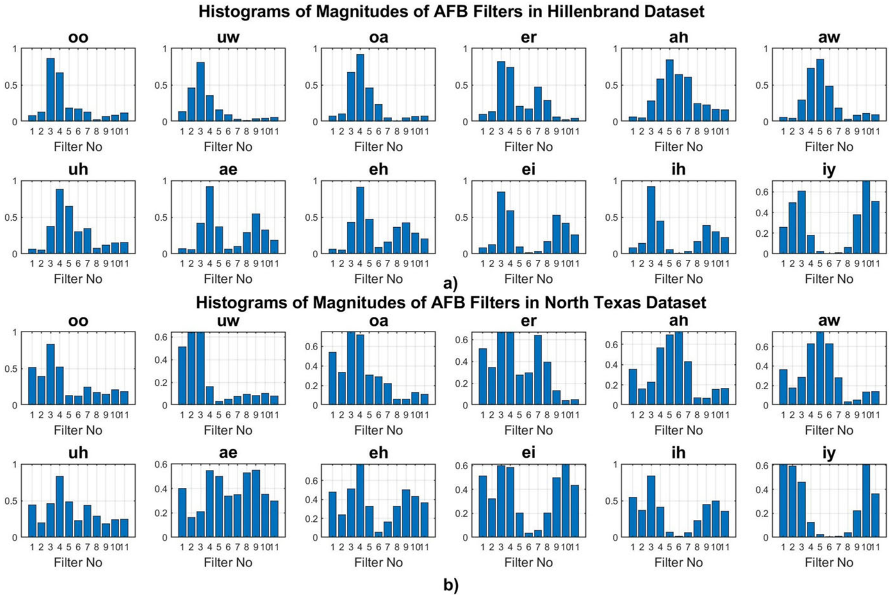


Fig. 8 Histograms of spectral magnitudes of AFB filters of vowels in the **a** Hillenbrand and **b** North Texas datasets. The horizontal axis denotes the AFB filter number

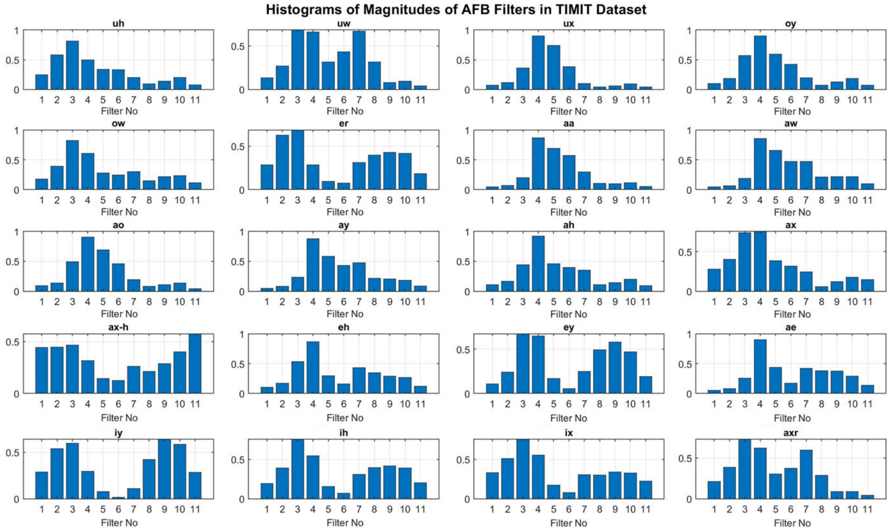


Fig. 9 Histograms of spectral magnitudes of AFB filters of vowels in the TIMIT dataset. The horizontal axis denotes the AFB filter number

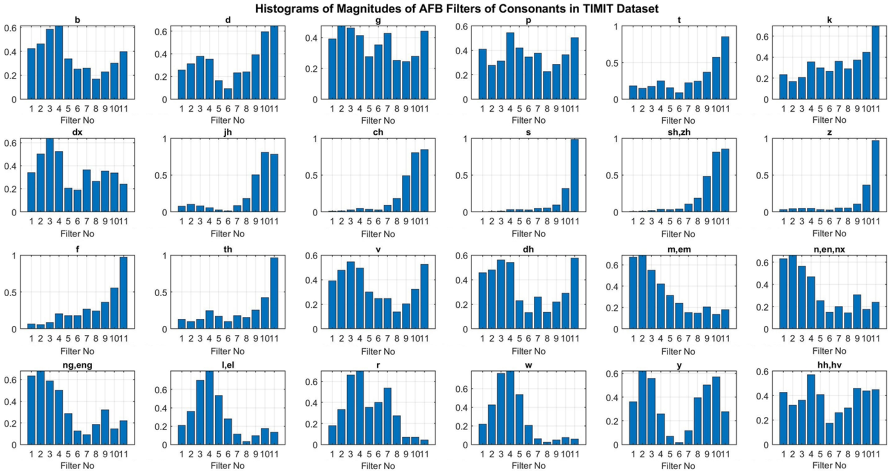


Fig. 10 Histograms of spectral magnitudes of AFB filters of consonants in the TIMIT dataset. The horizontal axis denotes the AFB filter number

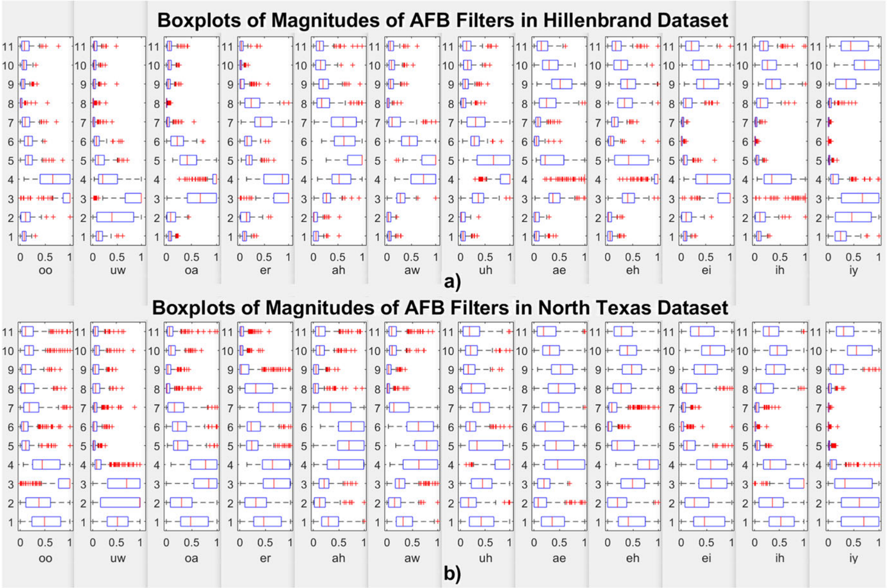


Fig. 11 Boxplots of the spectral magnitudes of the AFB filters in the **a** Hillenbrand and **b** North Texas vowel datasets. The vertical axis denotes the AFB filter number

that frequency regions do not help too much in consonants without detecting sharp changes during frequency formation.

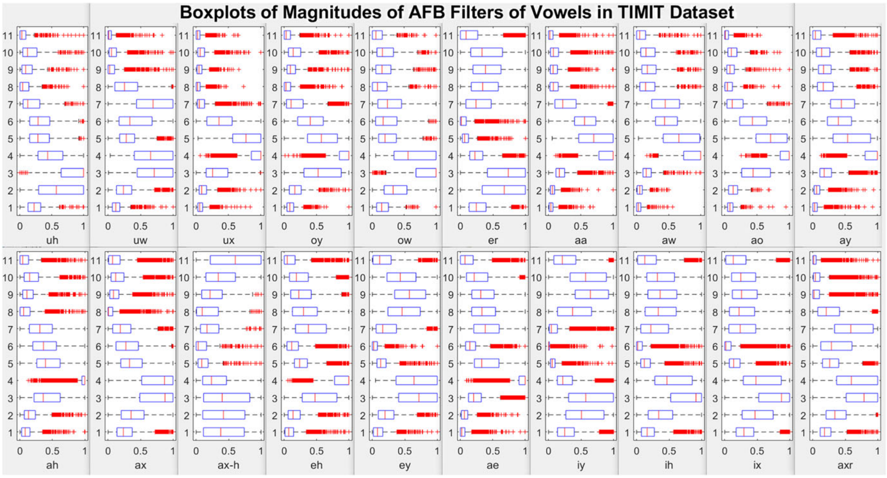


Fig. 12 Boxplots of the spectral magnitudes of the AFB filters of vowels in the TIMIT dataset. The vertical axis denotes the AFB filter number

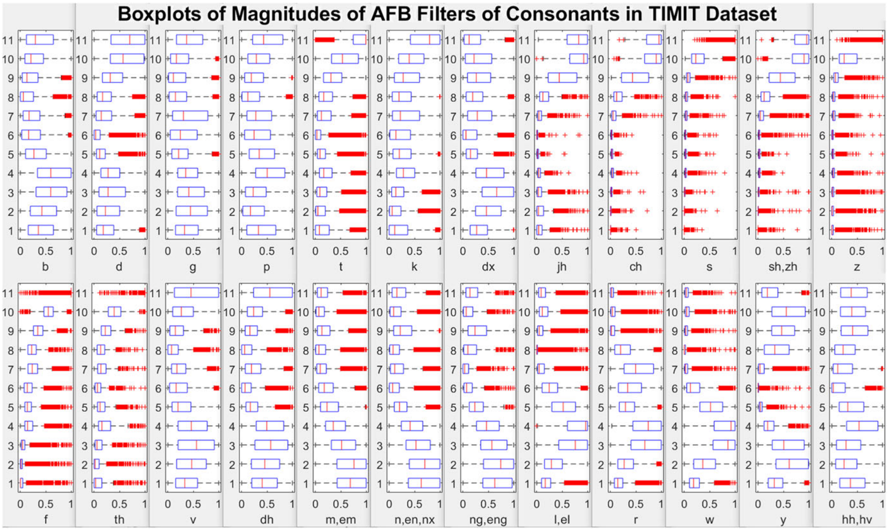


Fig. 13 Boxplots of spectral magnitudes of AFB filters of consonants in the TIMIT dataset. The vertical axis denotes the AFB filter number

4 Datasets

In this study, we used the Speech Command version 2 and the TIMIT datasets. The Speech Command version 2 contains 105,829 16-bit mono speech wave samples. Most of the samples are 1 s in length, the maximum length is 1 s, and the minimum length is 0.2133125 s; however, only 441 files are shorter than 0.5 s. Therefore, we set the sample length to 1 s and used zero padding when necessary. The total duration is 28.83 h. The second dataset is the widely used TIMIT dataset. The TIMIT dataset comprises sentences, words, and phones. In this study, we used phones only. There are 241,225 phones in the TIMIT dataset, including SA samples. The maximum phone length is 4.6428125 s, and the minimum phone length is 0.002 s. Note that in the TIMIT datasets, the silent parts are considered phones; otherwise, a single phone is not expected to be that long. The full duration of the data is 5.37 h. SCD consists of 35 different words and TIMIT comprises 39 phone classes as tabulated in Table 2.

In the TIMIT dataset, only 76 phones are longer than 1.015 s, 662 phones are longer than 0.511 s, and all are class 39 phones, which include silences and closures. There are 2590 phones longer than 0.25 s, 17,809 phones shorter than 0.0111 s and 1950 phones shorter than 0.0049 s. Therefore, we used a fixed sample point length of 4096 (~ 0.25 s) in the TIMIT dataset and padded with zero if needed. TIMIT contains 61 different phones; however, these 61 classes are mapped into 39 phones according to [69] in most of the classification applications, as depicted in Table 3. In Table 4 and Table 5, we represent the phones of the datasets used in this paper with their

Table 2 Number of samples in the Speech Command Dataset V2

Backward	1664	Five	4052	Learn	1575	One	3890	Tree	1759
Bed	2014	Follow	1579	Left	3801	Right	3778	Two	3880
Bird	2064	Forward	1557	Marvin	2100	Seven	3998	Up	3723
Cat	2031	Four	3728	Nine	3934	Sheila	2022	Visual	1592
Dog	2128	Go	3880	No	3941	Six	3860	Wow	2123
Down	3917	Happy	2054	Off	3745	Stop	3872	Yes	4044
Eight	3787	House	2113	On	3845	Three	3727	Zero	4052

Table 3 Number of phones in the TIMIT dataset (mapped from 61 to 39)

b	3067	th	1018	eh	5293
d	4793	v	2704	ey	3088
g	2772	dh	3879	ae	5404
p	3545	m, em	5600	aa, ao	8293
t	5899	n, en, nx	11,874	aw	945
k	6488	ng, eng	1787	ay	3242
dx	3649	l, el	9451	ah, ax, axh	8634
jh	1581	r	9064	oy	947
ch	1081	w	4379	ow	2913
s	10,114	y	2349	uh	756
sh, zh	3259	hh, hv	2836	uw, ux	3213
z	5046	iy	9663	er, axr	7636
f	3128	ih, ix	18,347	h#, bcl, dcl, epi, gcl, kcl, pcl, pau, q, tcl	53,488

Table 4 Vowels in the Hillenbrand, North Texas, and TIMIT datasets with ARPABET and IPA symbols

Samples	Hillenbrand	North Texas	TIMIT	IPA
heed	iy	ii	iy	/i/
hid	ih	il	ih, ix	/ɪ/
hayed, bait	ei	ee	ey	/e/
head	eh	el	eh	/ɛ/
had	ae	al	ae	/æ/
hod, pod	ah	aa	aa, ao	/a/
hawed, caught	aw	aw	aw	/ɔ/
hoed, boat	oa	oo	ow	/o/
hood	oo	ul	uh	/ʊ/
who'd, boot	uw	uu	uw, ux	/u/
hud, but	uh	uh	ah, ax, ax-h	/ʌ/
heard	er	er	er, axr	/ɝ/, /ɚ/
boy	–	–	oy	/ɔɪ/
bite	–	–	ay	/aɪ/

corresponding IPA and ARPABET symbols [44]. ARPABET is used to represent US English phones as distinct ASCII character pairs.

5 Convolutional Neural Networks

CNNs (convolutional neural networks) are very powerful and successful models for image recognition and pattern classification applications. Many models have been suggested for image recognition. Moreover, they are becoming increasingly popular in signal and speech processing applications and can perform even better than LSTM networks, which were designed for time series data naturally. The catch here is that time series data can be rearranged and fed into the classifier, similar to 2D image data. In speech processing applications, a signal is confined to a fixed frame in which it is assumed to be stationary. This transforms the problem into a standard image pattern matching problem. The advent of fast GPUs has enabled researchers to train and run CNN models faster than ever. In this study, we run our experiments with the famous Visual Geometry Group (VGG16) model [109]. VGG16 comprises thirteen convolutional and max pooling layers. At the end, 2 fully connected layers are connected to a softmax classifier. All convolutional layers are equipped with a rectified linear unit (ReLU) activation function [3] and batch normalization. The VGG16 model won the ILSVR (ImageNet Large-Scale Visual Recognition) image classification and localization challenge in 2014. It is an exceptionally large network and employs over 15 million parameters in our experiments. In the original VGG16, the input is fed into the model as 224 × 224 with 3 RGB channels. In this work, we arrange our 1-D speech signal data as 2-D data and send them to the model. We also remove the last 2 max

Table 5 All TIMIT phones with ARPABET and IPA symbols

Samples	TIMIT ARPABET	IPA	Samples	TIMIT ARPABET	IPA
beet, beat	iy	/i/	bee	b	b
bit	ih	/ɪ/	day	d	d
bait	ey	/eɪ/	gap	g	g
bet, met	eh	/ɛ/	pea	p	p
bat	ae	/æ/	tea	t	t
bott, bob, hod, cot	aa	/a/	key	k	k
bough, bout	aw	/aʊ/	muddy, dirty	dx	ɾ
boat	ow	/oʊ/	uh-oh, bat	q	ʔ
book, hood	uh	/ʊ/	joke, juice	jh	dʒ
boot	uw	/u/	choke, chair, cherry	ch	tʃ
butt, but	ah	/ʌ/	sea	s	s
bird	er	/ɝ/	she, shoe	sh	ʃ
about, the	ax	/ə/	zone	z	z
debit, roses, rabbit	ix	/ɨ/	azure, treasure, genre	zh	ʒ
about, story, bought, caught	ao	/ɔ/	fin	f	f
toot	ux	/ʉ/	thin, think, thick	th	θ
butter	axr	/ɚ/	van	v	v
suspect, potato	ax-h	/ə̥ /	then, this, those	dh	ð
boy	oy	/ɔɪ/	mom	m	m
bite	ay	/aɪ/	noon	n	n
obtain	bcl	b̚	sing	ng	ŋ
width	dcl	d̚	bottom	em	m̩
dogtooth	gcl	g̚	button	en	n̩
doctor	kcl	k̚	washington	eng	ŋ̍
accept	pcl	p̚	winner	nx	˜ɾ
catnip	tcl	t̚	lay	l	l
pause	pau	-	ray	r	r
epenthetic silence	epi	-	way	w	w
begin/end marker	h#	-	yacht	y	j
			hay, high	hh	h
			ahead	hv	ɦ
			bottle	el	l

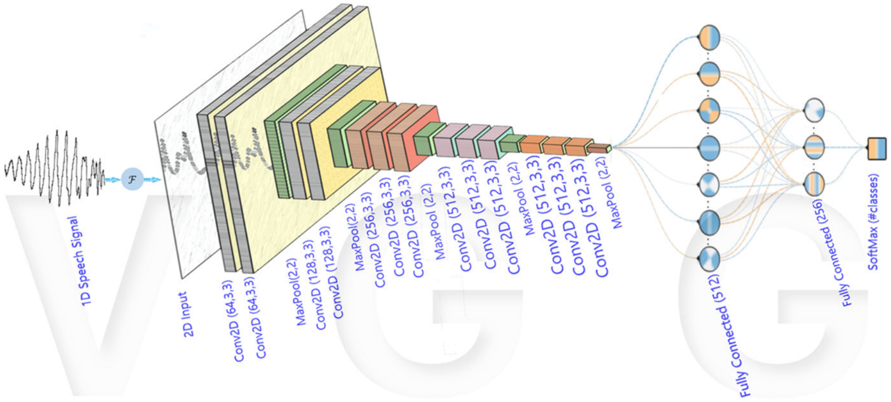


Fig. 14 VGG16 convolutional neural network model implementation

pooling layers for data structure compatibility. VGG16 is a great landmark in the quest to make computers understand what they see. The design of the implementation of our VGG16 model is shown in Fig. 14.

6 Results

In this section, we present the results of our experiments on the Speech Command V2 and TIMIT datasets with the VGG16 model. The experiments are conducted by employing AFB filters, Mel filters, PLP filters, and MFCC feature sets. The environmental setup is built on Python 3.8.10 [100], TensorFlow 2.11.0 [1], and Keras 2.11.0 [21]. Experiments are run using Adam optimization [62] with 0.001 learning rate, $\beta 1 = 0.9$, $\beta 2 = 0.999$, 100 epochs for the Speech Command dataset and 30 epochs for the TIMIT dataset. The data is split into 70% training and 30% test sets. The feature extraction phase is implemented in MATLAB 2019a [76] with the Auditory Toolbox [110]. PLP is implemented using the Rastamat package of Mark Shire and Dan Ellis [93]. Speech signals are dissected into 25-ms frames with 10-ms overlapping steps. We also incorporated the first-order delta acceleration coefficients in our feature sets. No data augmentation is performed on our datasets.

The feature extraction is run with 400 sample (25 ms) point frames and 160 (10 ms) sample point window shifts, thus creating a 24-frame 1D feature vector for TIMIT and 100-frame 1D feature vector for SCD dataset. This 1D feature vector is converted into 2D matrix form in the dimension of $(24 \times feature\_count(TIMIT), 100 \times feature\_count(SCD))$ while being fed into the VGG network. There are 22 features in the AFB filters, 26 in the MFCC filters, 42 in the PLP filters, and 80 in the Mel filters, including first-order delta acceleration coefficients. All speech signals are processed with Hamming window and pre-emphasis preprocessing with$\alpha = 0.97$. In the computations of the Mel and MFCC filters, the lowest frequency is 100 Hz, the linear spacing is 66 Hz, the number of linear filters is 13, the number of log filters is 27, and the log spacing is 1.0711703.

Table 6 Classification results (% ACC) on the Speech Command Dataset V2 with the VGG16 network. The number of features is shown in parentheses

Feature set	F score	Kappa	Accuracy (UA)
AFB filters (11)	95.46	95.30	95.45
Mel filters (40)	96.07	95.94	96.07
MFCC (13)	94.05	93.84	94.04
PLP (21)	95.23	95.07	95.22

Table 7 Classification results (% ACC) on the TIMIT dataset with the VGG16 network. The number of features is shown in parentheses

Feature set	F score	Kappa	Accuracy (UA)
AFB filters (11)	73.80	72.22	74.35
Mel filters (40)	77.05	75.24	77.04
MFCC (13)	73.09	70.95	73.09
PLP (21)	75.50	73.54	75.47

The classification results are tabulated in Table 6 for the Speech Command dataset and in Table 7 for the TIMIT dataset. AFB outperforms MFCC by a significant margin in both datasets and runs shoulder-by-shoulder with Mel and PLP filters in the Speech Command V2 dataset. AFB also converges better than does MFCC and strongly competes against Mel filters, as illustrated in the training accuracy graph of Fig. 15. AFB is also less susceptible to overfitting than are the MFCC, Mel, and PLP filters due to the smaller number of banks. We need to bear in mind that AFB filters contain only 11 coefficients compared to 40 Mel filters, 21 PLP filters, and 13 MFCCs.

We continue our quest to explore the best filter bank strategy and try various filter banks with different numbers of filters on the Speech Command V2 and TIMIT datasets with the VGG16 architecture. We attempt to increase the number of filters from 6 to 25 and introduce the results in Figs. 16 and 17. Selecting a filter count between 9 and 13 is a smart choice for implementing speech processing. The region below 9 (yellow) can be considered an underfitting area, and the region above 16 (red) is the area where overfitting concerns begin to emerge. Therefore, we have finalized our quest with 11 filters for the AFB filters to minimize the number of features.

7 Conclusions and Future Directions

In this study, we conducted experiments with novel AFB filters and compared them with Mel filters, MFCC and PLP features using the TIMIT dataset and the Speech Command Dataset version 2. The novel AFB filters always outperformed the MFCC in all the experiments and achieved accuracies comparable to those of the Mel filters in the Speech Command Dataset V2 when utilizing the famous VGG16 model. The results suggest that Mel filters, MFCC or PLP features can be replaced with novel AFB filters in speech processing applications. Using 40 banks in Mel filters seems unnecessary. The novel filter banks are computationally far less expensive than Mel

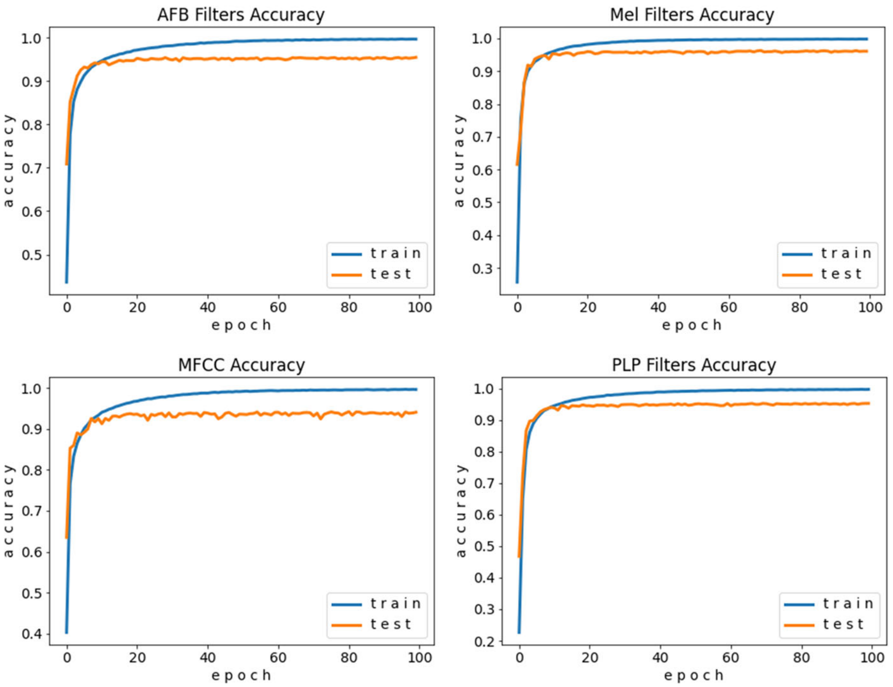


Fig. 15 Comparison of accuracy between AFB filters, Mel filters, and MFCC on the Speech Command Dataset V2 with VGG16 and the Adam optimizer. AFB converges better than MFCC and PLP and challenges Mel filters very strongly

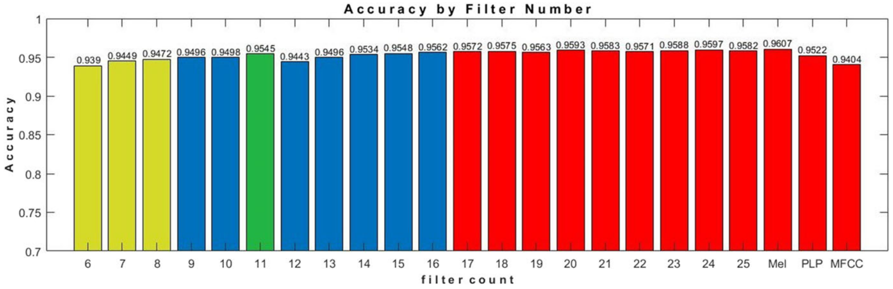


Fig. 16 Accuracies according to the number of filters on the Speech Command Dataset V2 with VGG16

filters. AFB contains only 11 filters compared to 40 Mel filters or 13 MFCCs and can be extracted faster. In this study, we evaluated different filter banks with up to 25 subbands. Some of these filter banks have nearly equal performances (filters with 17 and 19 subbands) with the Mel filters and can be used where high accuracy is the main objective. However, as in the Mel filters case, they can also be subject to overfitting concerns. We should also take into account the unbalanced nature of the TIMIT dataset, particularly regarding class 39. Our long-standing research points out that the number

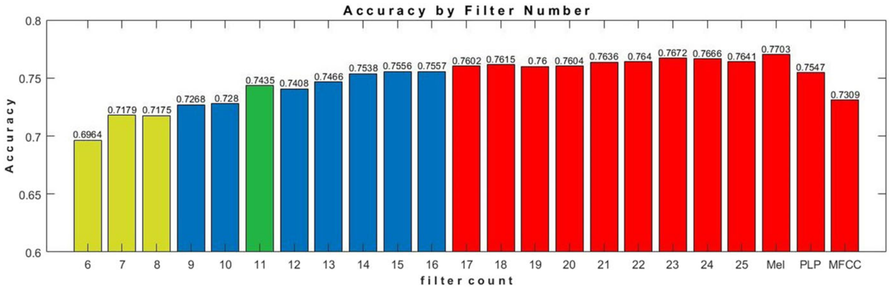


Fig. 17 Accuracies by the number of filters on the TIMIT dataset with VGG16. The region below 9 filters (yellow) points to underfitting, and the region above 16 filters (red) insinuates overfitting

of filters should be between 9 and 13; however, models with 17 or 19 subbands look like other strong candidates. As we discussed in Sect. 3, AFB filters enable filter bank summation methods and the use of wavelet filters such as *Fejér–Korovkin* or quadrature mirror filters for constructing nonuniform marginally overlapping trapezoidal filters, which will enable reconstruction of the signal using fast filter bank implementation algorithms. Moreover, AFB filters have proven to be powerful representations for speech processing due to their strong and natural foundation and may usher in new methods for speech processing applications. In our experiments, TIMIT is used as a database of phones, and the Speech Command dataset is used as a command dataset. This may be one of the effects of the results in the TIMIT dataset. Consonants cannot be segregated solely by frequency features, which is a well-established issue. If we can find a better representation for detecting voice onset time, phone boundaries, and consonant transitions, AFB filters may excel further. The performance of Mel filters as well as other filter banks with a large number of subbands in the TIMIT dataset is really intriguing and requires more sophisticated investigation. More research is needed here, particularly cross-corpora investigations to examine the generalization abilities of AFB filters, Mel filters, MFCC, and PLP or more accurate learning models may help. There is also a low possibility of creating more compact filter banks with fewer than 11 frequency bands. In future works, we will assess the effects of AFB filter banks on large vocabulary continuous speech recognition applications and other areas of speech processing, such as emotion recognition, speaker identification, gender detection, and speaker diarization, with advanced deep network models.

Author's Contributions C. Parlak: Conceptualization, methodology, writing, and experimentation. Y. Altun: Supervising, reviewing, writing, editing, validation.

Availability of Data and Material Data are available at the following site: https://github.com/cevparlak/AFB-Filters.

Declaration

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Consent to Participate and for Publication We confirm that we did not use participants in our study.

References

1. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, X. Zheng, et al. {TensorFlow}: a system for {large-scale} machine learning, in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16) Savannah Georgia USA* (2016), pp. 265–283
2. A.G. Adami, Automatic speech recognition: from the beginning to the Portuguese language, in *9th International Conference on Computational Processing of the Portuguese Language, Porto Alegre RS Brazil* (2010)
3. A.F. Agarap, Deep learning using rectified linear units (Relu), arXiv Preprint arXiv:1803.08375 (2018). https://doi.org/10.48550/arXiv.1803.08375
4. N. Ahmed, T. Natarajan, K.R. Rao, Discrete cosine transform. IEEE Trans. Comput. **C–23**(1), 90–93 (1974). https://doi.org/10.1109/T-C.1974.223784
5. E.P. Ahn, G.A. Levow, R.A. Wright, E. Chodroff, An Outlier analysis of vowel formants from a corpus phonetics pipeline, in *Proceedings of INTERSPEECH 2023* (2023), pp. 2573–2577. https://doi.org/10.21437/Interspeech.2023-1052
6. K. Allan (ed.), *The Oxford Handbook of the History of Linguistics* (OUP, Oxford, 2013). https://doi.org/10.1093/oxfordhb/9780199585847.001.0001
7. J. Allen, M.S. Hunnicutt, D.H. Klatt, R.C. Armstrong, D.B. Pisoni, *From Text to Speech: The MITalk System* (Cambridge University Press, Cambridge, 1987)
8. T. Arai, Sliding three-tube model as a simple educational tool for vowel production. Acoust. Sci. Technol. **27**(6), 384–388 (2006). https://doi.org/10.1250/ast.27.384
9. T. Arai, Education in acoustics and speech science using vocal-tract models. J. Acoust. Soc. Am. **131**(3), 2444–2454 (2012). https://doi.org/10.1121/1.3677245
10. E. Arısoy, L.M. Arslan, M.N. Demiralp, H.K. Ekenel, M. Kelepir, H.M. Meral, A.S. Özsoy, Ö. Şayli, O. Türk, B. Can-Yolcu, Duration of Turkish vowels revisited, in *12th International Conference on Turkish Linguistics (ICTL 2004) Dokuz Eylül Üniversitesi İzmir Türkiye* (2004), pp. 11–13
11. P.F. Assmann, W.F. Katz, Time-varying spectral change in the vowels of children and adults. J. Acoust. Soc. Am. **108**(4), 1856–1866 (2000). https://doi.org/10.1121/1.1289363
12. B.S. Atal, M.R. Schroeder, Adaptive predictive coding of speech signals. Bell Syst. Tech. J. **49**(8), 1973–1986 (1970). https://doi.org/10.1002/j.1538-7305.1970.tb04297.x
13. A. Berg, M. O'Connor, M.T. Cruz, Keyword transformer: A self-attention model for keyword spotting. arXiv preprint arXiv:2104.00769 (2021). https://doi.org/10.21437/Interspeech.2021-1286
14. J. Bernard, R. Mannell, A study of /h_d/ words in Australian English, in *Working Papers of the Speech, Hearing and Language Research Centre, Macquarie University* (1986)
15. G. Börtlü, The vowel triangle of Turkish and phonological processes of laxing and fronting in Turkish, (Master's Thesis) Hacettepe University (2020)
16. J.S. Bridle, M.D. Brown, An experimental automatic word-recognition system. JSRU Report No. 1003, Joint Speech Research Unit Ruislip England (1974)
17. K. Carki, P. Geutner, T. Schultz, Turkish LVCSR: towards better speech recognition for agglutinative languages, in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, Proceedings (Cat. No. 00CH37100)*, vol. 3 (2000), pp. 1563–1566. https://doi.org/10.1109/ICASSP.2000.861971
18. X. Chi, M. Sonderegger, Subglottal coupling and its influence on vowel formants. J. Acoust. Soc. Am. **122**(3), 1735–1745 (2007). https://doi.org/10.1121/1.2756793
19. E.R. Chodroff, M. Baese-Berk, Constraints on variability in the voice onset time of L2 English stop consonants, in *Proceedings of the 19th International Congress of Phonetic Sciences Melbourne, Australia* (2019). ISBN 978-0-646-80069-1
20. E. Chodroff, J. Godfrey, S. Khudanpur, C. Wilson, Structured variability in acoustic realization: a corpus study of voice onset time in American English stops, in *Proceedings of the 18th International Congress of Phonetic Sciences Glasgow, UK: the University of Glasgow* (2015). ISBN 978-0-85261-941-4
21. F. Chollet et al., Keras, GitHub. https://github.com/fchollet/keras. Accessed 1 Mar 2024

22. J. Coleman, J. Pierrehumbert, Stochastic phonological grammars and acceptability. arXiv preprint cmp-lg/9707017 (1997). https://doi.org/10.48550/arXiv.cmp-lg/9707017
23. S.A. Collins, Men's voices and women's choices. Anim. Behav. **60**(6), 773–780 (2000). https://doi.org/10.1006/anbe.2000.1523
24. F. Cox, An acoustic study of vowel variation in Australian English. (Doctoral dissertation, Macquarie University) (1996)
25. F. Cox, J. Fletcher, *Australian English Pronunciation and Transcription* (Cambridge University Press, Cambridge, 2017)
26. S. Dabbaghchian, Computational modeling of the vocal tract: applications to speech production. Doctoral dissertation, KTH Royal Institute of Technology Stockholm Sweden (2018)
27. S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 357–366 (1980). https://doi.org/10.1109/TASSP.1980.1163420
28. D.C. de Andrade, S. Leo, M.L.D.S. Viana, C. Bernkopf, A neural attention model for speech command recognition. arXiv preprint arXiv:1808.08929 (2018). https://doi.org/10.48550/arXiv.1808.08929
29. E. de Boer, Auditory physics. Physical principles in hearing theory. III. Phys. Rep. **203**, 125–231 (1991). https://doi.org/10.1016/0370-1573(91)90068-W
30. J.R. Deller, J.G. Proakis, J.H. Hansen, *Discrete-Time Processing of Speech Signals* (MacMillan Publishing Co, 2000). ISBN: 0-7803-5386-2
31. R.A. DePaolis, The intelligibility of words, sentences, and continuous discourse using the articulation index. J. Acoust. Soc. Am. **91**(6), 3584–3584 (1992). https://doi.org/10.1121/1.2029879
32. H. Diessel, Usage-based linguistics. Oxf. Res. Encycl. Linguist. (2017). https://doi.org/10.1093/acrefore/9780199384655.013.363
33. H. Dridi, K. Ouni, Towards robust combined deep architecture for speech recognition: experiments on TIMIT. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **11**(4), 525–534 (2020). https://doi.org/10.14569/IJACSA.2020.0110469
34. H. Dudley, R.R. Riesz, S.S. Watkins, A synthetic speaker. J. Frankl. Inst. **227**(6), 739–764 (1939). https://doi.org/10.1016/S0016-0032(39)90816-1
35. G. Fant, Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations (No. 2). Walter de Gruyter (1971). https://doi.org/10.1515/9783110873429
36. J.L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd edn. (Springer, New York, 2013). https://doi.org/10.1007/978-3-662-01562-9
37. H. Fletcher, W.A. Munson, Loudness, its definition, measurement, and calculation. J. Acoust. Soc. Am. **5**, 82–108 (1933). https://doi.org/10.1002/j.1538-7305.1933.tb00403.x
38. S. Fuchs, P. Birkholz, *Phonetics of Consonants. Oxford Research Encyclopedia of Linguistics* (Oxford University Press, Oxford, 2019). https://doi.org/10.1093/acrefore/9780199384655.013.410
39. J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, DARPA TIMIT acoustic-phonetic continuous speech corpus. LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium (1993). https://doi.org/10.35111/17gk-bn40
40. O. Ghitza, Robustness against noise: the role of timing-synchrony measurement, in *ICASSP '87 IEEE International Conference on Acoustics, Speech, and Signal Processing* (1987), pp. 2372–2375. https://doi.org/10.1109/ICASSP.1987.1169917
41. J. Goldsmith, B. Laks, Generative phonology: its origins, its principles, and its successors, *The Cambridge History of Linguistics* (2006). https://doi.org/10.13140/RG.2.2.29518.25923
42. A. Haar, Zur Theorie der orthogonalen Funktionensysteme. Math. Ann. **69**(3), 331–371 (1910). https://doi.org/10.1007/BF01456326
43. R.E. Hagiwara, *Acoustic Realizations of American /r/ as Produced by Women and Men*. University of California Los Angeles (1995)
44. A.K. Halberstadt, Heterogeneous acoustic measurements and multiple classifiers for speech recognition (Doctoral dissertation, Massachusetts Institute of Technology), (1999)
45. S.M. Harding, G.F. Meyer, Formant continuity and auditory scene analysis: the effect of vowel formant manipulations on the perception of synthetic nasal consonants. J. Acoust. Soc. Am. **109**(5), 2312–2312 (2001). https://doi.org/10.1121/1.4744120
46. S. Harding, G. Meyer, Changes in the perception of synthetic nasal consonants as a result of vowel formant manipulations. Speech Commun. **39**(3–4), 173–189 (2003). https://doi.org/10.1016/S0167-6393(02)00014-6
47. B. Hayes, *Introductory Phonology* (Wiley, New York, 2008)

48. S. Herculano-Houzel, The human brain in numbers: a linearly scaled-up primate brain. Front. Hum. Neurosci. (2009). https://doi.org/10.3389/neuro.09.031.2009
49. H. Hermansky, Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. **87**(4), 1738–1752 (1990). https://doi.org/10.1121/1.399423
50. H. Hermansky, N. Morgan, A. Bayya, P. Kohn, RASTA-PLP speech analysis, in *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, vol. 1 (1991, December), pp. 121–124. https://doi.org/10.1109/ICASSP.1992.225957
51. J. Hillenbrand, R.T. Gayvert, Vowel classification based on fundamental frequency and formant frequencies. J. Speech Lang. Hear. Res. **36**(4), 694–700 (1993). https://doi.org/10.1044/jshr.3604.694
52. J. Hillenbrand, L.A. Getty, M.J. Clark, K. Wheeler, Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. **97**(5), 3099–3111 (1995). https://doi.org/10.1121/1.411872
53. J.M. Hillenbrand, M.J. Clark, C.A. Baer, Perception of sinewave vowels. J. Acoust. Soc. Am. **129**(6), 3991–4000 (2011). https://doi.org/10.1121/1.3573980
54. W. Holmes, *Speech Synthesis and Recognition* (CRC Press, Boca Raton, 2002)
55. M. Huckvale, Exploiting speech knowledge in neural nets for recognition. Speech Commun. **9**(1), 1–13 (1990). https://doi.org/10.1016/0167-6393(90)90040-G
56. G. Hunter, H. Kebede, Formant frequencies of British English vowels produced by native speakers of Farsi, in *Acoustics* (2012)
57. P.L.M. Johannesma, The pre-response stimulus ensemble of neurons in the cochlear nucleus, in *IPO Symposium on Hearing Theory, Eindhoven Netherlands* (1972), pp. 58–69
58. K. Johnson, K. Johnson, Acoustic and auditory phonetics. Phonetica **61**(1), 56–58 (2004). https://doi.org/10.1159/000078663
59. D. Jurafsky, J.H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing. Computational Linguistics, and Speech Recognition* (2000)
60. D. Kewley-Port, Y. Zheng, Vowel formant discrimination in ordinary listening conditions I. J. Acoust. Soc. Am. **100**(4_Supplement), 2689–2689 (1996). https://doi.org/10.1121/1.417026
61. D. Kewley-Port, Y. Zheng, Vowel formant discrimination: Towards more ordinary listening conditions. J. Acoust. Soc. Am. **106**, 2945–2958 (1999). https://doi.org/10.1121/1.428134
62. P.D. Kingma, J. Ba. "Adam: a method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014). https://doi.org/10.48550/arXiv.1412.6980
63. G. Kidd, C.R. Mason, V.M. Richards, F.J. Gallun, N.I. Durlach, W.A. Yost, R.R. Fay, *Auditory Perception of Sound Sources* (Springer, New York, 2008), pp.143–189
64. R. Kirchner, Turkish vowel harmony and disharmony: an Optimality theoretic account, in *Rutgers Optimality Workshop I 22* (1993, October), pp. 1–20
65. D.H. Klatt, Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Am. **67**(3), 971–995 (1980). https://doi.org/10.1121/1.383940
66. A. Klautau, Classification of Peterson & Barney's vowels using Weka. Federal University of Para Brazil: Technical report (2002)
67. P. Ladefoged, K. Johnson, *A Course in Phonetics*, 7th edn. (Cengage Learning, USA, 2014). ISBN 10: 1285463404 ISBN 13: 978128546340
68. W. Lawrence, The synthesis of speech from signals which have a low information rate. W. Jackson editor Communication Theory Butterworths Sci. Pub. London, 460–469 (1953)
69. K.F. Lee, H.W. Hon, Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Acoust. Speech Signal Process. **37**(11), 1641–1648 (1989). https://doi.org/10.1109/29.46546
70. S.V. Levi, Glides, Laterals, and Turkish vowel harmony (Master's thesis, University of Washington), (2000)
71. X. Li, Z. Zhou, Speech command recognition with convolutional neural network. CS229 Stanford Education 31 (2017)
72. A.M. Liberman, K.S. Harris, H.S. Hoffman, B.C. Griffith, The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. **54**, 358–368 (1957). https://doi.org/10.1037/h0044417
73. R. Lyon, A computational model of filtering, detection, and compression in the cochlea, in *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing* (1982), pp. 1282–1285. https://doi.org/10.1109/ICASSP.1982.1171644
74. E. Maeda, N. Usuki, T. Arai, N. Saika, Y. Murahara, Comparing the characteristics of the plate and cylinder type vocal tract models. Acoust. Sci. Technol. **25**(1), 64–65 (2004). https://doi.org/10.1250/ast.25.64

75. M. Malik, M.K. Malik, K. Mehmood, I. Makhdoom, Automatic speech recognition: a survey. Multimed. Tools Appl. **80**, 9411–9457 (2021). https://doi.org/10.1007/s11042-020-10073-7
76. The Math Works, Inc., MATLAB (Version 2019a), [Computer software], https://www.mathworks.com/. Accessed 1 March 2024
77. K. Migimatsu, I.T. Tokuda, Experimental study on nonlinear source–filter interaction using synthetic vocal fold models. J. Acoust. Soc. Am. **146**(2), 983–997 (2019). https://doi.org/10.1121/1.5120618
78. A. Mittal, M. Dua, Automatic speaker verification systems and spoof detection techniques: review and analysis. Int. J. Speech Technol. (2022). https://doi.org/10.1007/s10772-021-09876-2
79. M.R. Molis, Perception of vowel quality in the F2/F3 plane. The University of Texas at Austin (2002)
80. H. Møller, C.S. Pedersen, Hearing at low and infrasonic frequencies. Noise Health **6**(23), 37–57 (2004)
81. T. Nguyen, Total number of synapses in the adult human neocortex. Undergrad. J. Math. Model. One+Two **3**(1), 26 (2010). https://doi.org/10.5038/2326-3652.3.1.26
82. M. Nielsen, On the construction and frequency localization of finite orthogonal quadrature filters. J. Approx. Theory **108**(1), 36–52 (2001). https://doi.org/10.1006/jath.2000.3514
83. D.Z. Obidovna, Distinctive features of male and female oral speech in modern English. Int. J. Lit. Lang. **2**(10), 14–21 (2022)
84. W. O'Grady, M. Dobrovolsky, F. Katamba (eds.), *Contemporary Linguistics* (St. Martin's, New York, 1997)
85. G.S. Ohm, Über die definition des tones, nebst daran geknüpfter theorie der sirene und ähnlicher tonbildender vorrichtungen. Ann. Phys. Chem. **59**, 513–565 (1843)
86. H.F. Olson, *Music, Physics and Engineering* (Dover Publications. 1967), pp. 248–251. ISBN 978-0-486-21769-7
87. A.V. Oppenheim, *Discrete-Time Signal Processing* (Pearson Education India, 1999)
88. F. Orság, Speaker dependent coefficients for speaker recognition. Int. J. Secur. Appl. **4**(1), 31–34 (2010)
89. P. Padmini, D. Gupta, M. Zakariah, Y.A. Alotaibi, K. Bhowmick, A simple speech production system based on formant estimation of a tongue articulatory system using human tongue orientation. IEEE Access **9**, 4688–4710 (2020). https://doi.org/10.1109/ACCESS.2020.3048076
90. G.E. Peterson, H.L. Barney, Control methods used in a study of the vowels. J. Acoust. Soc. Am. **24**(2), 175–184 (1952). https://doi.org/10.1121/1.1906875
91. C. Phillips, K. Govindarajan, A. Marantz, D. Poeppel, T. Roberts, H. Rowley, E. Yellin, MEG studies of vowel processing in auditory cortex. *Poster presented at Cognitive Neuroscience Society meeting Boston* (1997)
92. J. Picone, Fundamentals of speech recognition: A short course. Institute for Signal and Information Processing, Mississippi State University (1996)
93. The PLP and RASTA in MATLAB, [Computer Software], https://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/, Accessed 1 Mar 2024
94. L. Rabiner, B.-H. Juang, *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, 1993)
95. L.R. Rabiner, R.W. Schafer, Introduction to digital speech processing. Found. Trends® Signal Process. **1**(1–2), 1–194 (2007). https://doi.org/10.1561/2000000001
96. L. Rabiner, R. Schafer, *Theory and Applications of Digital Speech Processing* (Prentice Hall Press, Englewood Cliffs, 2010)
97. H. Reetz, A. Jongman, *Phonetics: Transcription, Production, Acoustics, and Perception* (Wiley, New York, 2020)
98. D.W. Robinson, R.S. Dadson, A re-determination of the equal-loudness relations for pure tones. Br. J. Appl. Phys. **7**, 166–181 (1956). https://doi.org/10.1088/0508-3443/7/5/302
99. G. Rosen, Dynamic analog speech synthesizer. J. Acoust. Soc. Am. **30**, 201–209 (1958). https://doi.org/10.1121/1.1909541
100. G.V. Rossum, F.L. Drake, *Python 3 Reference Manual* (CreateSpace, Scotts Valley, 2009)
101. N. Saika, E. Maeda, N. Usuki, T. Arai, Y. Murahara, Developing mechanical models of the human vocal tract for education in speech science, in *Proceedings of the 2002 Forum Acusticum Sevilla Spain* (2002)
102. H.A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, M. Rosa-Zurera, Age group classification and gender recognition from speech with temporal convolutional neural networks. Multimed. Tools Appl. **81**(3), 3535–3552 (2022). https://doi.org/10.1007/s11042-021-11614-4

103. J.F. Schouten, The residue revisited, in *International Symposium on Frequency Analysis and Periodicity Detection in Hearing, June 23–27, 1969, Driebergen, The Netherlands, Sijthoff* (1970), pp. 41–58
104. M.R. Schroeder, *Computer Speech: Recognition, Compression, Synthesis*, vol. 35 (Springer, New York, 2004)
105. A. Seebeck, Beobachtungen über einige bedingungen der entstehung von tönen. Ann. Phys. Chem. **53**, 417–436 (1841)
106. A. Sek, B.C. Moore, Frequency discrimination as a function of frequency, measured in several ways. J. Acoust. Soc. Am. **97**(4), 2479–2486 (1995). https://doi.org/10.1121/1.411968
107. S. Seneff, A joint synchrony/mean-rate model of auditory speech processing. J. Phon. **16**(1), 55–76 (1988). https://doi.org/10.1016/S0095-4470(19)30466-8
108. M. Siafarikas, I. Mporas, T. Ganchev, N. Fakotakis, Speech recognition using wavelet packet. J. Wavel. Theory Appl. **2**(1), 41–59 (2008)
109. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014). https://doi.org/10.48550/arXiv.1409.1556
110. M. Slaney, Auditory toolbox. Interval Research Corporation, Tech. Rep, **10**(1998), 1194 (1998)
111. S.S. Stevens, J. Volkmann, E.B. Newman, A scale for the measurement of the psychological magnitude pitch. J. Acoust. Soc. Am. **8**(3), 185–190 (1937). https://doi.org/10.1121/1.1915893
112. K.N. Stevens, *Acoustic Phonetics* (MIT Press, Cambridge, 1998)
113. C. Stilp, E. Chodroff, "Please say what this word is": Linguistic experience and acoustic context interact in vowel categorization. JASA Express Lett. **3**(8), 085203 (2023). https://doi.org/10.1121/10.0020558
114. Y. Suzuki, H. Takeshima, Equal-loudness-level contours for pure tones. J. Acoust. Soc. Am. **116**(2), 918–933 (2004). https://doi.org/10.1121/1.1763601
115. L. Tóth, Phone recognition with hierarchical convolutional deep maxout networks. EURASIP J. Audio Speech Music Process **2015**(1), 1–13 (2015). https://doi.org/10.1186/s13636-015-0068-3
116. V.A. Trinh, H.S. Kavaki, M.I. Mandel, Importantaug: a data augmentation agent for speech, in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (8592–8596), IEEE* (2022). https://doi.org/10.1109/ICASSP43922.2022.9747003
117. R.S. Turner, The Ohm–Seebeck dispute, Hermann von Helmholtz, and the origins of physiological acoustics. Br. J. Hist. Sci. **10**(1), 1–24 (1977). https://doi.org/10.1017/S0007087400015089
118. N. Umeda, Linguistic rules for text-to-speech synthesis. Proc. IEEE **64**(4), 443–451 (1976). https://doi.org/10.1109/PROC.1976.10153
119. P.P. Vaidyanathan, *Multirate Systems and Filter Banks* (Pearson Education India, Delhi, 2006)
120. P. Warden, Speech Commands: A dataset for limited-vocabulary speech recognition (2018). arXiv preprint arXiv:1804.03209. https://doi.org/10.48550/arXiv.1804.03209
121. J.G. Wells, A study of the formants of the pure vowels of British English (Doctoral dissertation, University of London) (1962)
122. M. Wereski, The threshold of hearing. STEAM J. **2**(1), 20 (2015). https://doi.org/10.5642/steam.20150201.20
123. I. Wilson, Using Praat and Moodle for teaching segmental and suprasegmental pronunciation, in *Proceedings of the 3rd International WorldCALL Conference: Using Technologies for Language Learning (WorldCALL 2008)* (2008), pp. 112–115
124. D. Woods, E.W. Yund, T.J. Herron, M.A. Cruadhlaoich, Consonant identification in consonant-vowel-consonant syllables in speech-spectrum noise. J. Acoust. Soc. Am. **127**(3), 1609–1623 (2010). https://doi.org/10.1121/1.3293005
125. W.A. Yost, Pitch perception. Atten. Percept. Psychophys. **71**(8), 1701–1715 (2009). https://doi.org/10.3758/APP.71.8.1701
126. E.C. Zsiga, *The Sounds of Language: An Introduction to Phonetics and Phonology* (Wiley, New York, 2024)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.