Chapter 8

Temporal Filterbanks in Cochlear Implant Hearing and Deep Learning Simulations

Payton Lin

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/66627

Abstract

The masking phenomenon has been used to investigate cochlear excitation patterns and has even motivated audio coding formats for compression and speech processing. For example, cochlear implants rely on masking estimates to filter incoming sound signals onto an array. Historically, the critical band theory has been the mainstay of psychoacoustic theory. However, masked threshold shifts in cochlear implant users show a discrepancy between the observed critical bandwidths, suggesting separate roles for place location and temporal firing patterns. In this chapter, we will compare discrimination tasks in the spectral domain (e.g., power spectrum models) and the temporal domain (e.g., temporal envelope) to introduce new concepts such as profile analysis, temporal critical bands, and transition bandwidths. These recent findings violate the fundamental assumptions of the critical band theory and could explain why the masking curves of cochlear implant users display spatial and temporal characteristics that are quite unlike that of acoustic stimulation. To provide further insight, we also describe a novel analytic tool based on deep neural networks. This deep learning system can simulate many aspects of the auditory system, and will be used to compute the efficiency of spectral filterbanks (referred to as "FBANK") and temporal filterbanks (referred to as "TBANK").

Keywords: auditory masking, cochlear implants, filter bandwidths, filterbanks, deep neural networks, deep learning, machine learning, compression, audio coding, speech pattern recognition, profile analysis, temporal critical bands, transition bandwidths

1. Introduction

The transformation of sound into a representation within the auditory system involves many layers of information analysis and processing. Sound is first converted into nervous impulses by cochlear hair cells, which are mechanically organized to distribute the spectral energy of their excitation along the length of the basilar membrane (**Figure 1**). The connecting nerve fibers show a bandpass response to the input signal, where the density of firings for a particular fiber varies with the stimulus intensity over a certain range. Basic information from a sound is then extracted and passed to subsequent stages for perceptual machinery to present its own construction of reality. Noninvasive methods are needed to investigate the influence of these higher-level perceptual processes on the properties of the cochlea. For example, the method of psychophysical inference is often used to fill in the gaps of physiological knowledge.

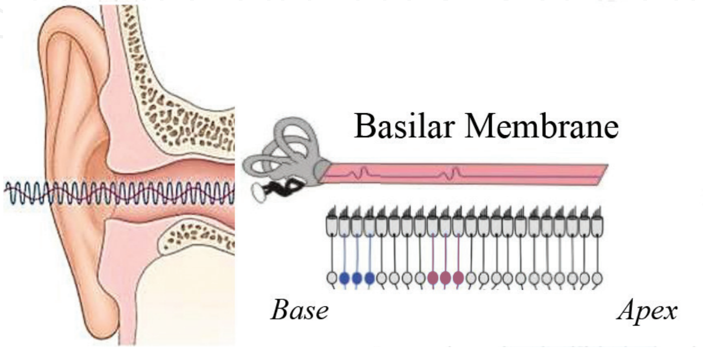


Figure 1. Spatial arrangement of cochlear hair cells along the basilar membrane (base to apex).

This chapter will be divided into two sections. In the section on human hearing research, we will predict neural firing characteristics from the perspective of psychoacoustic "*masking*" experiments. In the section on machine hearing research, we will compare artificial neural network input from the perspective of machine learning experiments. Experimental data is presented in both human hearing and machine hearing to supplement the incompleteness of current neurophysical methods by providing new insight into the stages of processing.

2. Human hearing research

2.1. Auditory masking

2.1.1. Spectral masking

The masking phenomenon of one tone by another provides quantitative data on frequency selectivity and the dynamical theory of the cochlea [1]. In a psychoacoustic experiment, the testing stimulus is called the probe, the sound that interferes with the detection of the probe is called the masker, and the amount of masking refers to the amount by which the hearing threshold of the probe is raised in the presence of the masker. The method of measuring a threshold shift is straightforward. First, the detection threshold of the probe is determined. Next, the threshold shift of the probe is determined in the presence of the masker. Auditory masking curves have established wide-ranging mathematical relationships between sensory behavioral responses and even the activity of single neurons [2]. In general, the probe is most easily masked by sounds with frequency components that are close to the probe.

2.1.2. Critical band theory

The "*critical band theory*" [3] has played an important role in hypothesizing how the auditory system resolves the components of complex sound. In classic band-widening experiments, the threshold of a sinusoidal probe is measured as a function of the bandwidth of a noise masker. First, a noise with a constant power density is centered at the probe frequency. As the bandwidth increases, the total noise power increases (which presumably has effects on the threshold for detecting the probe).

Masking curves have shown that the threshold of the probe increases at first, but flattens off as the addition of more noise (at a greater distance from the probe frequency) produces no additional masking. The bandwidth at which the probe threshold ceases to increase is called the "*critical bandwidth*." To account for these observations, the listener is assumed to make use of a filter with a center frequency close to the probe when detecting the probe in noise. According to this critical bandwidth theory, the noises outside the range of the filter should presumably have no effect on detection. If this filter passes the signal and removes much of the noise, then only the components of the noise that passes through the filter should have any effect in masking the probe. Therefore, thresholds should correspond to a certain signal-to-noise ratio at the output of the auditory filter.

According to this *power spectrum model* of masking, all stimuli are represented by their long-term power spectra (or the relative phases of the components) while short-term fluctuations in the masker are ignored. **Figure 2** shows the typical estimates of energy detection. Although *energy detection models* remain fundamental to theories of auditory perception, the axiom that energy only passes by a single auditory filter has been contradicted multiple times [4–7]. These findings violate the fundamental assumption of critical band theory and therefore challenge previous estimates of peripheral filtering.

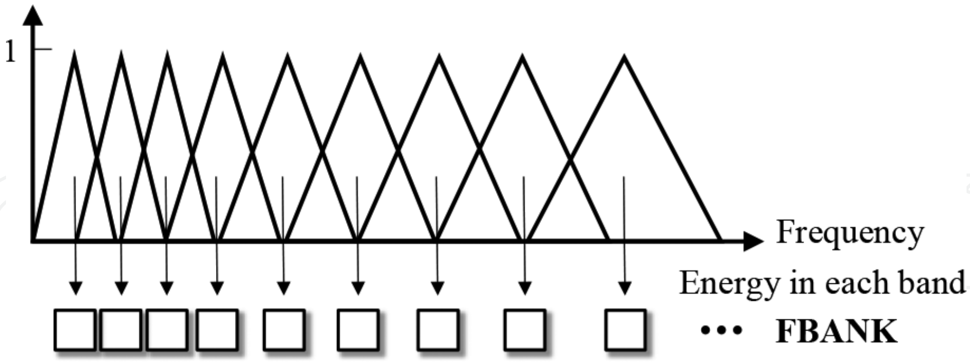


Figure 2. Energy detector model where the basilar membrane behaves as if it contains a bank of bandpass filters with overlapping passbands, where each point along the basilar membrane corresponds to a filter with a different center frequency. This Fourier-transform-based log filterbank with spectral coefficients distributed on a mel-scale is often referred to as "FBANK" in audio-coding applications involving machine learning. In the section on machine hearing research, the computational efficiency of FBANK will be evaluated in deep learning systems.

2.1.3. Profile analysis

According to the critical bandwidth theory, a tone added to noise should be detected by an increase in the energy from a single auditory filter centered at the signal frequency. On the contrary, experimental manipulations (e.g., roving-level procedures) that degrade energy cues in tone-in-noise detection tasks show no effects on detection thresholds. Listeners must therefore rely on alternative cues instead of just spectral analysis of a stimulus to explain the data of level-invariant detection (where single-channel energy cues are severely disrupted). In "*profile analysis*" [5], this process is described as detecting changes in the overall shape of the spectrum. With across-channel cues, listeners are able to compare the shape or profile of the outputs of different auditory filters to enhance signal detection.

2.1.4. Temporal critical bands

Temporal discrimination tasks offer an alternative to spectral critical bands. In a temporal model, the detection cues are thought to be temporal in nature and based on changes in the cadence of neural discharge. These temporal models contrast with energy detection models that assume a rate-place neural code. Recently, auditory filter bandwidths were measured for a temporal process using an amplitude-modulation (AM) detection task [6]. The critical bandwidth for a temporal process (referred to as "*temporal critical band*") was observed to be consistently greater than that predicted by the critical band theory. Therefore, these findings decrease confidence in previous estimates of peripheral filtering.

2.1.5. Transition bandwidths

Discontinuous threshold functions also contradict spectral critical bandwidths by implying that the discrimination tasks evoke different and separate auditory processes. For instance, "*transition bandwidths*" [7] assume that envelope cues dominate at narrow bandwidths, while across-channel level comparisons dominate at wide bandwidths. This concept stresses that there are changes in the underlying process, unlike the constraining boundaries of a solitary process (as hypothesized in critical bandwidth or energy integration theories). For transition bandwidths, the changes to another dominant auditory process are thought to be due to a central mechanism (whereas critical bandwidths are only associated with the periphery). Therefore, transition bandwidths allow for multiple filtering processes to occur.

2.1.6. The volley theory

The divergence of positions between spectral bandwidths and temporal bandwidths shares similar controversies as the *place theory* and the *temporal theory* of pitch perception. The place theory states that the perception of sound depends on where each component of frequency produces vibrations. The temporal theory states that the perception of sound depends on the temporal patterns of neurons responding to sound in the cochlea. The "*volley theory*" [8] postulates that groups of neurons in the auditory system respond to firing action potentials that are slightly out-of-phase with one another so that they can be combined to encode and send a greater frequency of sound to the brain for analysis, as shown in **Figure 3**. In the next sections, we describe the importance of resolving these theories and assumptions to improve real-world solutions for data *compression* and speech processing. For instance, we will compare the efficiency of systems that use only one filterbank or multiple filterbanks.

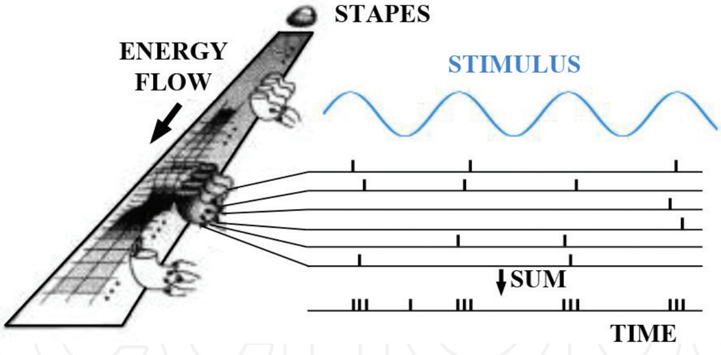


Figure 3. Temporal properties of nerve firing according to the "volley theory" of temporal coding.

2.2. Filterbanks

2.2.1. Audio coding and data compression

Audio coding formats that use lossy data compression take advantage of human auditory masking properties [9]. For instance, the MP3 format hides noises under the signal spectrum based on the masking property that sounds near the threshold of another sound will either be completely masked or reduced in loudness. These auditory masking properties also play critical roles in both speech coding applications and objective quality measures. In the next section, we will cover the impact of auditory masking on the coding of *cochlear implants* (devices that require data compression due to the electroneural bottleneck).

2.2.2. Cochlear implants

The cochlear implant is a surgically implanted electronic device that restores partial hearing to a person who is deaf or hearing impaired [10–12]. This neural prosthesis provides similar functions of the inner ear by electrically stimulating the auditory nerve. Cochlear implants consist of an external microphone, a speech processor, a transmitter, an internal receiver, and a multielectrode array stimulator. The microphone is placed on the ear and picks up incoming sounds from the environment. The speech processor filters these incoming sounds into different frequency channels and sets the appropriate electrical stimulation parameters. Next, a transmitter coil powers and transmits the processed sound signals through the skin to the internal receiver. Finally, the receiver converts the signals into electric impulses that stimulate an electrode array. Electrode arrays are surgically coiled within the scala tympani of the cochlea so that individual electrode plates can electrically stimulate different regions of the auditory nerve. A sparse electric representation is sufficient for the restoration of hearing.

Cochlear implant performance is satisfactory in quiet settings, but the abnormal perception of electric pitch limits the performance in noise. Typical users only detect over a 10% change in pitch compared to normal-hearing listeners who easily detect <1% change. Electric pitch is degraded because only a limited number of electrodes (~22 electrodes) can be inserted into the cochlea versus the >3000 inner hair cell transducers in a normal cochlea. In addition, the current spread from each electrode is uncontrollably broad and large areas of nerves can be unintentionally activated. Spectral mismatches can also occur from degraded nerve survival or inaccurate frequency-to-electrode allocation [11].

Speech recognition in noise becomes especially difficult without the ability to adequately separate components of sound from interfering sources. **Figure 4** shows the cochlear implant coding scheme [12] that was discovered to greatly improve speech recognition. This sparse representation at the auditory periphery is unique as it presents electric pulse stimulations that feature both (1) temporal envelope information and (2) place information. Section 3.2.3 will discuss how this coding is simulated as temporal envelope bank (TBANK) features in deep learning systems [13].

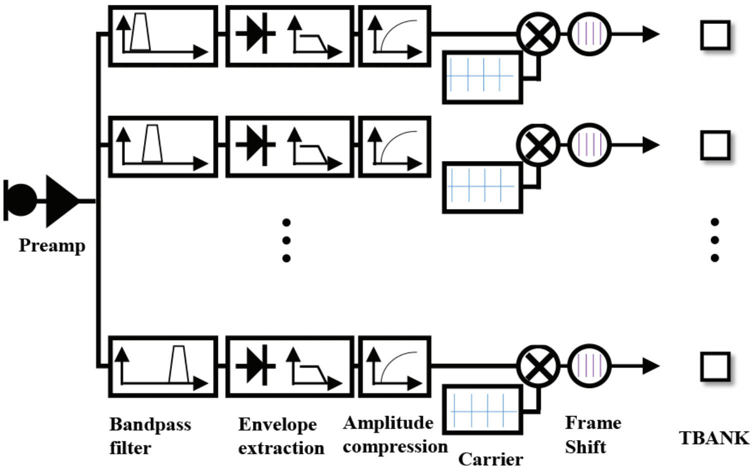


Figure 4. Temporal properties of a cochlear implant processor. In machine learning, temporal features can be derived from extracted temporal envelope bank (referred to as "TBANK").

2.3. Auditory masking in cochlear implants

To optimize current settings, psychoacoustic experiments were designed to investigate how the human auditory system processes complex sound interactions from electric stimulations. Specifically, auditory masking was investigated using electric stimulations as the probe or masker to understand how electric stimulations separate into individual sound sources. The diversified subject population with different types of hearing loss and electric configurations also provide alternative testing paradigms to reevaluate previous masking results obtained from normal hearing subjects. By measuring electric stimulations, the research field can gain new insight to study the interactions of peripheral and central auditory systems. In this section, we will review previous comparisons of ipsilateral *electric-on-electric* masking, *electric-on-acoustic* masking, and also contralateral *electric-on-electric* masking. We will then compare auditory masking curves in cochlear implants with the recently proposed concepts of profile analysis, temporal critical band, and transition bandwidths in normal hearing.

2.3.1. Comparison of electric-on-electric masking

Similar to the observations in normal hearing, electric masking studies [14] have shown that the amount of forward masking increases by decreasing the spatial separation between the probe electrodes and the electric pulses from adjacent maskers. Both the amount of masking and the spread of neural excitation increase with electric masker levels. Cochlear implant excitation patterns were also shown to have a spatial bandpass characteristic with a peak in the region of the masked electrode [15].

2.3.2. Comparison of electric-on-acoustic masking

Electric-on-acoustic masking can also be measured for cochlear implant users who have preserved residual acoustic hearing following implantations (**Figure 5**). In a unilateral cochlear implant user with functional hearing preserved in the implanted ear, electric stimulations were observed to interact in the peripheral and central auditory system [16]. The masking growth function in **Figure 5** shows the detection thresholds of an electrode increased when the level of a 125-Hz acoustic masker increased from 90 to 110 dB. The 250-Hz acoustic masker also elevated electric detection in a similar manner. This data is consistent with the central theory of auditory masking and even provides new supporting evidence since the acoustic stimulations had to have been confined to the functional hair cells or nerves (as there is no known mechanism that acoustic stimulations could have directly activated a nerve fiber).

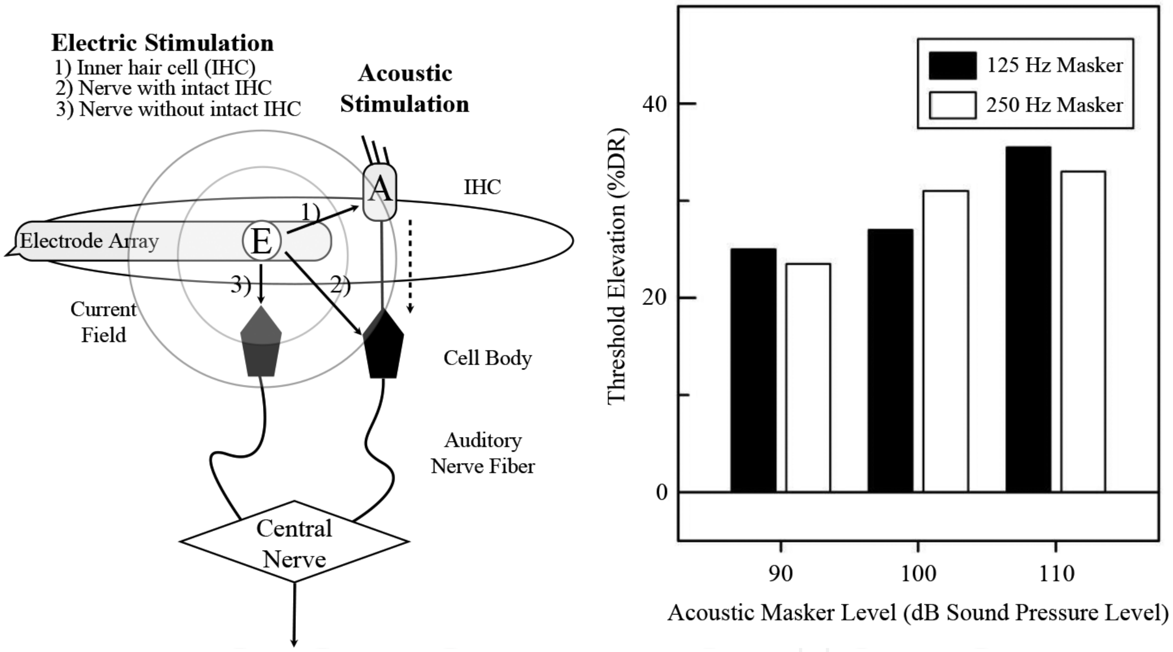


Figure 5. Ipsilateral masking data from [16]. The left panel shows a schematic representation of the acoustic (A) masking and the electric (E) masking mechanisms in hybrid hearing. The right panel shows the masking of an electrode probe by acoustic maskers at 125 or 250 Hz.

2.3.3. Comparison of contralateral electric-on-electric masking

Contralateral electric-on-electric masking can also be measured in bilateral cochlear implant users [17]. **Figure 6** shows the complete set of central masking data, with threshold elevation normalized so that each function peaks at 1. Each of the bilateral subjects was tested twice, alternating the ear used as the masker or probe (n = 14). As shown in **Figure 6**, the contralateral masking electrodes elevated the detection thresholds in both the left and the

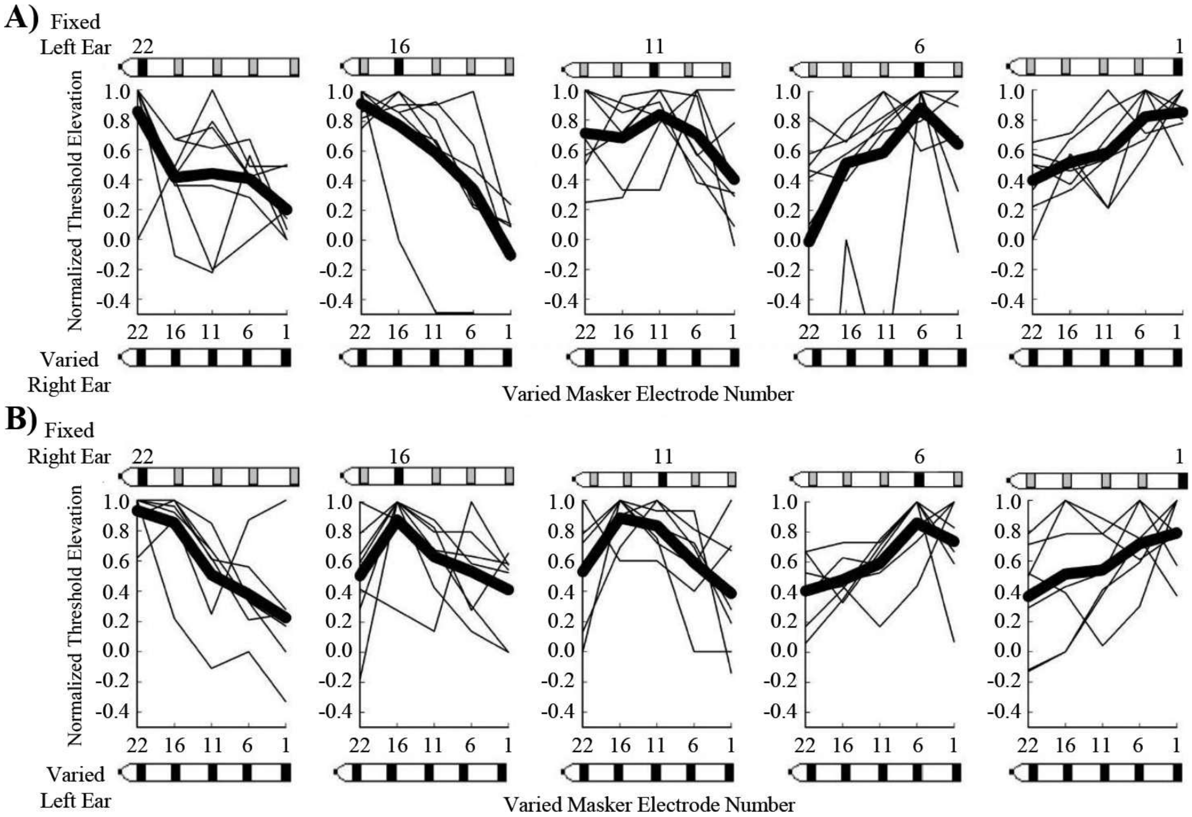


Figure 6. Central masking data measured and replotted for seven bilateral cochlear implant subjects from [17]. The curves are sorted into panels according to the location of the probe electrode number (black contacts) in either the (A) fixed left ear or (B) fixed right ear. Thin lines show the individual central masking curves and the thick lines show the mean data for each fixed probe electrode location. All curves are normalized so that the peak threshold is equal to 1.

right ears. The threshold elevation peaks generally occurred between interaural pairings sharing the same electrode number (which corresponds to electrodes with similar insertion depths across ears).

Figure 7 presents the same data to show the growth of masking as a function of the masker-probe electrode separation across ears. Masker-probe separation is calculated by subtracting the differences between the masker and probe electrode numbers. Place-matched masking conditions are categorized as "0" since both the masker and probe electrode numbers were identical across ears. When categorized in this manner, **Figure 7** shows the amount of central masking diminished with masker-probe electrode separation. In [17], this data was reorganized to analyze the growth of masking. A two-way repeated measure analysis of variance (ANOVA) showed a significant main effect of masker-probe electrode separation and threshold elevation [$F(2.122, 27.581) = 3.667$, $p = 0.036$]. There was also a significant main effect of masker-probe electrode separation, ear used as the probe electrode, and threshold elevation [$F(2.563, 33.323 = 9.472$, $p < 0.001$]. The masking growth pattern for each ear was also fitted with exponential equations and displayed similar spatial constants and significant R^2 values ($R^2 > 0.97$). The results demonstrate that the amount of central masking diminished with masker-probe electrode separation at similar rates on both sides.

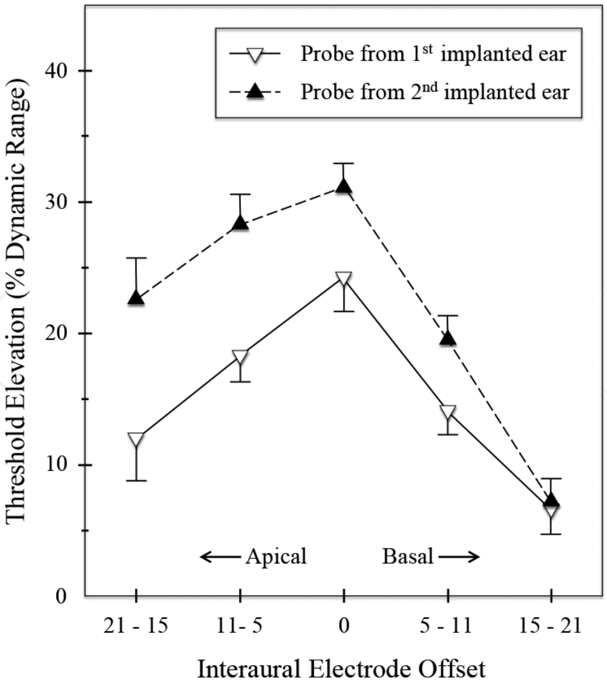


Figure 7. Threshold elevation as a function of the interaural electrode offset between the masker and the probe (*n* = 14). "Apical" refers to all masking conditions where the masker was apically positioned from the probe, whereas "basal" refers to all conditions where the masker was basally positioned from the probe. The 1st and 2nd implanted ears refer to the ears on each side of a participating subject that were sequentially implanted in two separate surgeries.

Figures 6 and **7** show bilateral cochlear implant stimulation contralaterally masked in a place-dependent manner. For electric-on-electric signals, the average thresholds peaked when the position of the masker and probe electrodes were place-matched across ears and diminished with electrode separation. These place-dependent findings have also been reconfirmed in a recent work [18] using different electrode arrays and testing apparatus. However, both studies [17–18] directly counter a previous conclusion in [19] that central masking with bilateral users gives rise to increased threshold, but not in a place-dependent manner (as is the case for contralateral masking in normal hearing). This previously accepted hypothesis in [19] was most likely concluded from data that was obscured by a limited test population (*n* = 2), electrical malfunctions (arrays with multiple electrical shorts), subjects reporting discomfort (uncomfortably high-pitched sensations), and electrodes that were inserted with primitive surgical techniques (offsets in electrode place accuracy of 6–9 electrodes).

2.3.4. Comparison of masking in cochlear implants and normal hearing

It will be important to optimize the configurations of cochlear implant stimulation as it has been reported that electrical pulse trains and acoustic sine waves do not fuse or merge well into a single percept [17, 20]. The presence of electric masking has indicated regions where electric and acoustic signals share similar frequencies as normal acoustic masking. However, the data in **Figure 6** show a large amount of individual variability as many of the bilateral participants displayed unmatched masking patterns across ears. This individual variability could be the result of several factors. First, cochlear implant users are likely to have irregular patterns of auditory nerve survival, which could cause the current to stimulate auditory fibers that are too apical or too basal from the intended location. “Dead regions” in auditory nerves may also prevent the adjacent masking electrodes from stimulating distinct place-frequencies. Second, different surgical procedures could result in variability between the insertion depths of a user's electrode array. This variability could be significant since neural activation patterns depend on the density of neurons in a particular region of the cochlea and on the radial distance between the electrode array and neural targets in the modiolus.

In general, cochlear implant subjects have exhibited electric-masking patterns that are much broader compared to what has been observed in normal hearing [14–20]. A reduction in the magnitude of contralateral versus ipsilateral masking functions was observed in [18], but this reduction was not as great as observed in normal hearing. Together, these findings of broader masking with cochlear implants both support and are supported by the concepts of:

1. Distorted *profile analysis*, where cochlear implant users are unable to adequately use across-channel cues to compare the shape of the output of different auditory filters.
2. Reliance on *temporal critical bands*, where cochlear implant users relied on filter bandwidths that were consistently broader than predicted by critical band theory.
3. Rapidly changing *transition bandwidths*, where cochlear implant users used separate and different auditory processes to handle either the narrow or wide bandwidths.
4. Distorted *volley theory* of encoding, where cochlear implant users were unable to combine the phases of action potentials to analyze a greater frequency of sound.

These findings all decrease confidence in previous estimates of peripheral filtering as well as the assumptions made in *critical band theory*, especially when combined with recent evidence in normal-hearing listeners that suggest a flexible selection of spectral regions upon which to base across-frequency comparisons [21]. Furthermore, the wide bandwidths observed in the initial filters of the cochlear implant subjects directly contradicts the theory that extraction of envelope information should be constrained to a single auditory filter, as theorized in [22]. For these reasons, *transition bandwidths* [7, 23] are the most plausible solution as they explain patterns that were observed in both normal hearing and electric hearing experiments (as this concept allows an interplay to occur between temporal or spectral processes). In the section on machine hearing research, a novel method based on “*deep learning*” is utilized to prove the computational efficiency of transition bandwidths in artificial neural network systems.

3. Machine hearing research

3.1. Motivation to compare human and machine hearing systems

There are several factors that have confounded cochlear implant research. Psychoacoustic experiments are often rendered inconclusive due to large individual variabilities in cochlear implant subjects. The physical limitations of uncontrollable test populations include variable nerve survival before implantation, inter-implant intervals and usage time, neuroplasticity after implantation, age at testing, and surgical insertion depths. These physical limitations have significant effects on the amount of masking [16, 17]. In addition, individual variability can arise due to cognitive factors or subjective testing protocols. For instance, the evaluation of cochlear implants can be unintentionally influenced by decision rules or dynamic ranges based on loudness judgment, visual feedback, sequential test order, unrealistic simulations, and even the content material used in subjective studies such as speech recognition [24, 25]. Therefore, alternative methods such as computational simulations and mathematical models should be used in order to account for these uncontrollable factors of individual variation.

We can only appreciate how sophisticated the human auditory system truly is when trying to simulate perceptual processing on a computer. By building a computational model, we gain new insight and develop quantitative ways to analyze each step of signal processing. Computational models are well suited to investigate how information from independent fibers are distributed and the extent to which distinct bandpass filterbanks are constructed within neural architectures [4]. In this chapter, we use artificial neural networks in order to measure response properties of auditory fibers using realistic representations of different integrative processes. Machine learning can provide algorithms for understanding learning in neural systems and can even benefit from these ongoing biological studies [26].

3.2. Deep neural networks (DNNs)

3.2.1. Automatic speech recognition: an auditory perspective

There have been many attempts to incorporate principles of human hearing into machine systems [27]. The motivation for these previous attempts was simply that human perception is much more stable than machines over a range of sources of variability. Therefore, it was reasonable to expect that the functional modeling of the human subsystems could provide plausible direction for machine research. One of the first auditory-inspired features (**Figure 2**) was based on the mel-scale warping of the spectral frequency axis (referred to as “FBANK”), which is then parameterized as mel-frequency cepstral coefficients (referred to as “MFCC”) [28]. The usual objective for selecting an appropriate representation is to compress the input data by eliminating the information that is not pertinent for analysis and to enhance those aspects of the signal that contributes significantly to the detection of differences. In *automatic speech recognition* (ASR), these MFCC features were shown to allow better suppression of insignificant spectral variation in the higher-frequency bands. Concatenating other types of auditory-inspired spectro-temporal features with MFCCs can also boost performance [29]. In [30], cochlear implant speech synthesized from subband temporal envelope was shown to contain sufficient information to rival MFCC features in terms of accuracy. These acoustic simulations of cochlear implants [31] were subsequently proposed as general indicators to conduct useful subjective studies. In [32], the cross-disciplinary methods of cognitive science and machine learning were converged to promote the shared views of computational [33] foundations. Our study [32] expanded on [30] by comparing cochlear implant results using the Bayesian model of human concept learning [34] and proposed hidden Markov models (HMMs) for computationally predicting cochlear implant performance. In the next sections, we will further expand upon previous studies by introducing state-of-the-art tools based on *deep neural network* algorithms and presenting new results comparing the efficiency of profile analysis, temporal critical band, and transition bandwidths in cochlear implant simulations.

3.2.2. Spectral filterbank (FBANK) features as input to deep learning systems

Deep neural networks (DNNs) (**Figure 8**) make use of gradient-based optimization algorithms to adjust parameters throughout a multilayered network based on the errors at its input [35]. In DNNs, multiple processing layers learn representations of data with multiple levels of abstractions. In deep hierarchal structures, the internal layers of DNNs provide learned representations of the input data. The benefit of studying filterbank learning in DNNs is that the filterbank input can be viewed as an extra layer of the network, where these filterbank parameters are updated along with the parameters in subsequent layers [36, 37].

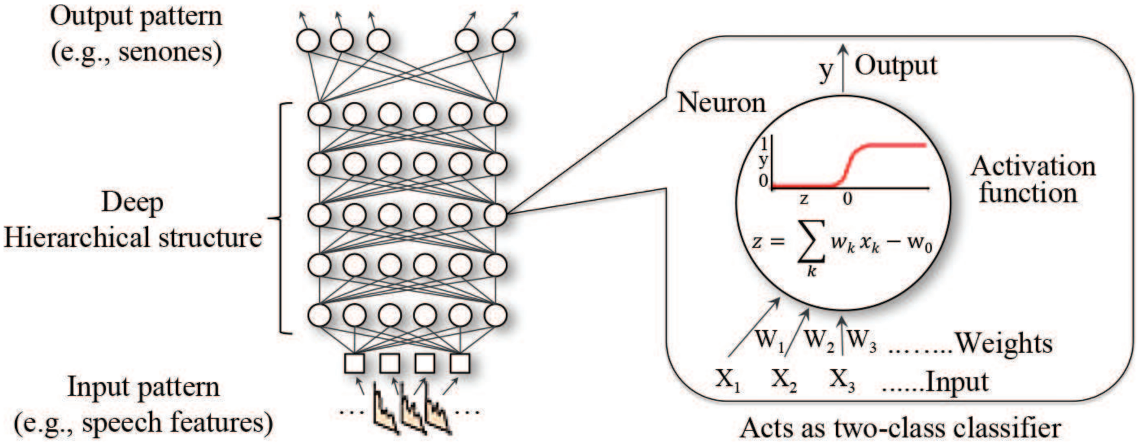


Figure 8. Structure of a deep neural network (DNN). Deep learning allows multiple layers of nonlinear processing.

In machine learning, speech is viewed as a two-dimensional signal where the spatial and temporal dimensions have vastly different characteristics. For instance, the time-dynamic information in the high-frequency regions is different compared to low-frequency regions. Although FBANK is popular, Sainath et al. [36, 37] argued that features designed based on the critical band theory might not guarantee appropriate frameworks for the end goal of reducing error rates. Since the power-spectra removes information from the signal by computing from a fixed window-length, FBANK features often lack the necessary temporal information. By starting with a raw signal representation to learn filterbanks jointly in a DNN framework, the results computed in [36] share many similarities as concepts from psychoacoustic studies:

1. Consistent with *critical band theory*, the computational results showed a similarity between learned and mel-filters in the low-frequency regions.

2. Consistent with *transition bandwidths*, the computational results showed the learned filters had multiple peaks in the mid-frequency regions (indicating that multiple important critical frequencies are being picked up, rather than just one like the mel).

3. Consistent with the *volley theory*, the computational results showed the learned filters are high-pass filters compared to mel-filters (which are bandpass at high-frequency regions).

3.2.3. Temporal envelope bank (TBANK) features

Our work in [13, 32] derived an alternative input feature (**Figure 4**) for ASR based on temporal envelope bank (referred to as "TBANK") which was inspired by *temporal critical bands* [6] and the broad temporal masking patterns of cochlear implants [16, 17]. The TBANK features have been evaluated as an input feature for DNNs [38] and as a temporal alignment feature for DNNs [39]. In the present study, we will combine both FBANK and TBANK features to improve the temporal dimension and its correlation with the frequency or spatial-domain properties in DNNs. **Figure 9** shows FBANK+TBANK (referred to as □□, "double-BANK") features, which were inspired by psychoacoustic results showing the flexible usage of across-channel cues [21], transition bandwidths [7, 23], and the volley theory [8].

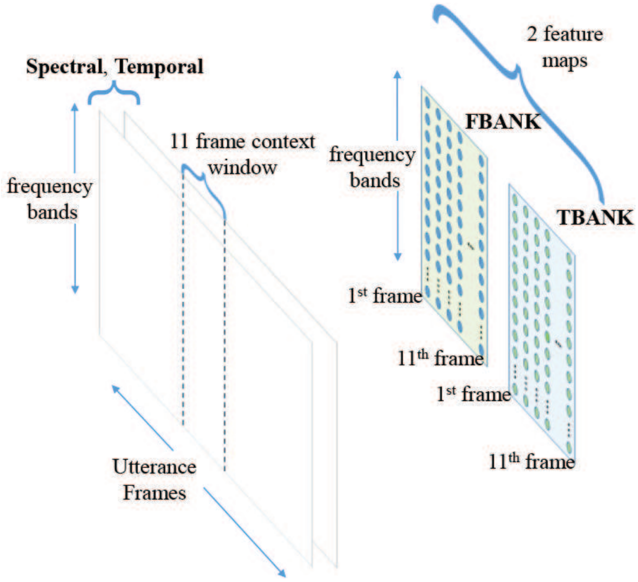


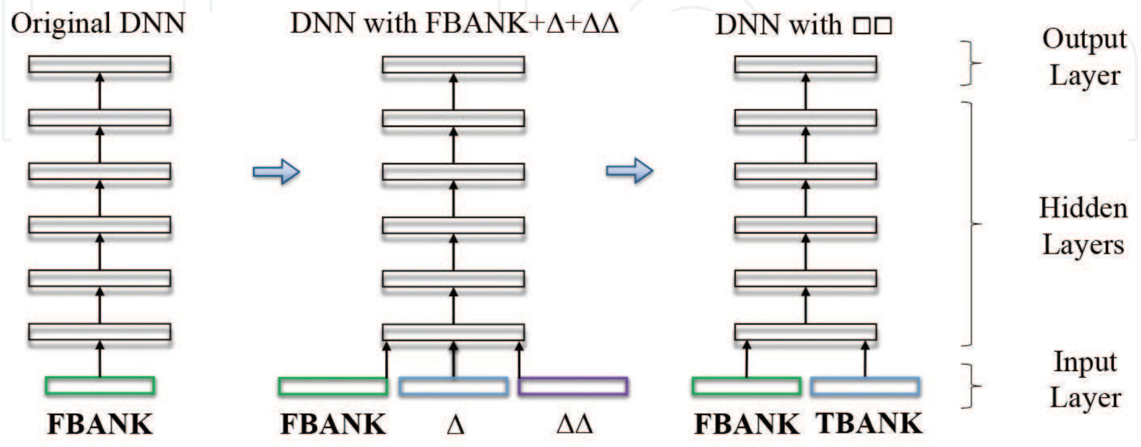
Figure 9. A simplified FBANK+TBANK (referred to as □□, "*double-BANK*") representation of speech.

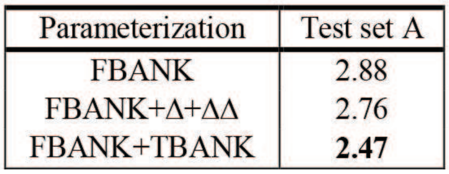
In Section 3.3, we will use the same procedures in [38, 39] for the Aurora-4 robustness task (with a cochlear implant speech processor available at: www.tigerspeech.com/angelsim).

3.3. Computational results

3.3.1. Comparison of FBANK and TBANK on a computational ASR task

"Raw" TBANK features were derived from 32 channels of band envelope (**Figure 4**) via white-noise carriers [38]. These features were designed to preserve temporal and amplitude cues in each spectral band, but remove the spectral detail within each band as explained in [12]. Δ and ΔΔ dynamic features were computed from derivative values with respect to time [40]. In **Table 1**, context-dependent DNN-HMMs were trained using 40-dimensional FBANK (described in [36]), 120-dimensional FBANK+Δ+ΔΔ (described in [41]), and our 80-dimensional □□ (FBANK+TBANK) input representation. It should be noted that the computational cost of changing the size of the input layer is negligible. **Table 1** shows inclusion of TBANK in the □□ (FBANK+TBANK) input features yielded a 14% improvement compared to FBANK and an 11% improvement over the FBANK+Δ+ΔΔ representation.





Parameterization	Test set A
FBANK	2.88
FBANK+Δ+ΔΔ	2.76
FBANK+TBANK	**2.47**

Note: **Bold** indicates better score.

Table 1. DNN performance (error rate %) on clean training set.

3.3.2. Comparison of FBANK and TBANK on temporal alignment task

Table 2 shows error rates for *Gaussian mixture model* (GMM)-HMMs when trained and tested on TBANK (**Figure 4**) alignment features (**Figure 10**) with white-noise carrier [13, 39]. TBANK models models aligned the training data to create senone labels for training the DNN. The results in **Table 2** show the temporally aligned DNN gives fewer errors when subsequently trained and tested on FBANK features.

Tree-building features	**Error % (GMM)**	**Error % (DNN)**
MFCC	5.08	2.88
16 band envelopes	5.44	**2.80**
24 band envelopes	**5.03**	**2.63**
32 band envelopes	5.90	**2.82**

Note: **Bold** indicates better score.

Table 2. Comparison of different tree-building features to generate a state-level alignment on the training set. TBANK features had 16, 24, or 32 band envelopes via white-noise carrier.

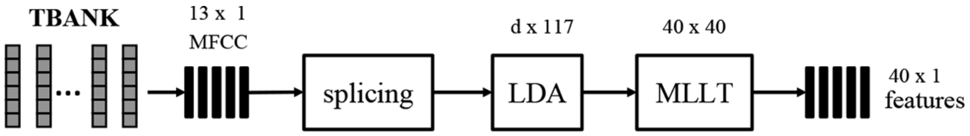


Figure 10. Generation of temporal alignment features using extracted TBANK, as in [39].

Designing a representation that both preserves relevant detail in the speech signal and also provides stability/invariance to distortions is a nontrivial task. Therefore, **Figure 11** derives slowly varying amplitude modulation (AM) and frequency modulation (FM) from speech to design novel features (referred to as frequency amplitude modulation encoding “FAME”) with different modulations, as proposed in [42] and computed in ASR [31, 39]. In the FAME condition, the FM is smoothed in terms of both rate and depth and then modulated by the AM. The “slow” FM tracks gradual changes around a fixed frequency in the subband. The FAME stimuli are obtained by additionally frequency modulating each of the band's center frequency before amplitude modulation and subband summation. Finally, FAME stimuli were used to derive alternative features via extracted TBANK (**Figure 10**) for tree building and temporal alignment in GMM systems.

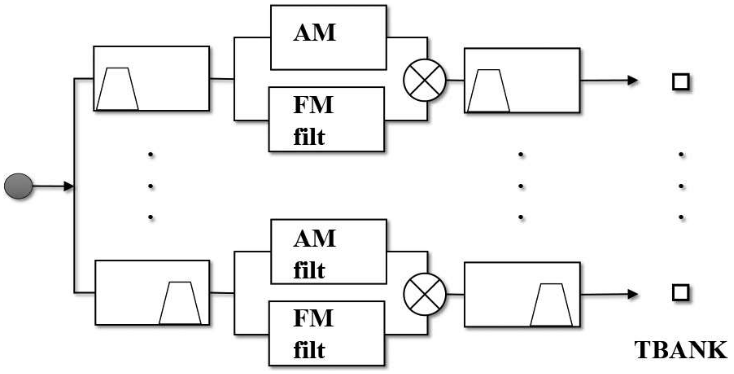


Figure 11. Signal processing diagram of *frequency amplitude modulation encoding* (FAME).

Compared to band envelope features (**Figure 4**), **Table 3** shows AM and FM lowered the error rate in GMM systems used during forced alignment to generate frame-level DNN training labels. These results expand upon [31] and provide additional evidence that FAME preserves more of the relevant detail compared to other carriers.

Tree-building TBANK features	Error rate % (GMM-Alignment)
16 band envelopes	5.44
16 bands of FAME	**5.21**

Note: **Bold** indicates better score.

Table 3. Comparison of alternative TBANK features for GMM alignment systems.

Table 4 shows error rates for GMM systems trained and tested on an additive configuration of □□ (MFCC + FAME) (**Figure 12**). This configuration was inspired by a frequency-dependent model that explains the loudness function in human auditory systems [43]. In this two-stage model, the first stage of processing is performed by a mechanical mechanism in the cochlear (for high-frequency stimuli) and by a neural mechanism in the cochlear nucleus (for low-frequency stimuli). **Table 4** shows the DNN gives fewer errors during time alignment with this additive configuration of □□ (MFCC + FAME) when subsequently trained and tested on FBANK. By digitally adding the different high-frequency FAME information (via TBANK) to the low-frequency MFCC information, **Table 4** shows this additive □□ (MFCC + FAME) feature representation allowed a better alignment in GMM systems during the generation of senone training labels for training the DNN.

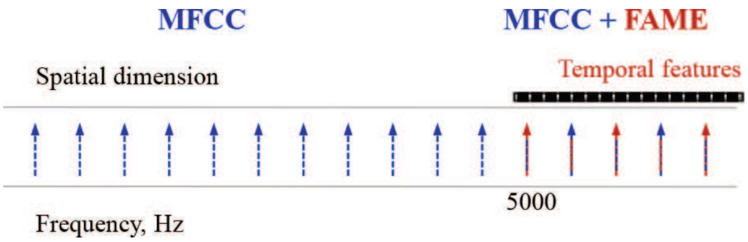


Figure 12. A digitally additive configuration of □□ alignment features (MFCC + FAME).

Table 5 provides further analysis [39] of deletion, substitution, or insertion errors to quantify the effects of the digitally additive □□ (MFCC + FAME) configuration. Misclassification leads to substitution errors. An imperfect segmentation leads to deletion errors (when some sounds are completely missed) or insertion errors (from extra boundaries). By reducing the extra segment boundaries, **Table 5** shows how front-end FAME processing using the FM extraction at high-frequency regions solves the segregating and binding problems [42] in ASR systems. The □□ results also demonstrate the computational efficiency of multiple filterbanks, which supports temporal critical bands [6] and transition bandwidths [7, 23].

Tree-building features	WER% (GMM)	WER% (DNN)
MFCC	5.08	2.88
MFCC + FAME	**4.82**	**2.75**

Note: **Bold** indicates better score.

Table 4. MFCC vs. □□ alignment feature (three additive FAME bands at high-frequency regions).

Tree-building features	(GMM) Deletion, substitution, insertion	(DNN) Deletion, substitution, insertion
MFCC	19, 189, 64	13, 114, 27
MFCC + FAME	20, **177**, **61**	18, **101**, 28

Note: **Bold** indicates less errors.

Table 5. Error type (deletion, substitution, insertion) analysis.

3.3.3. Discussion relating human hearing with machine hearing research

Many computational algorithms [9, 27, 28, 29, 30, 31, 35, 40, 41, 42] have been inspired by auditory processing pathways in the human nervous system. Traditionally, the critical band theory is commonly accepted for baseline DNN features [35] due to the ability of MFCC and FBANK to allow better suppression of insignificant spectral variation in the higher-frequency bands. However, recent progress in deep learning systems has allowed computational models that are composed of multiple layers of parallel processing to learn representations of filterbank features with multiple levels of abstraction. In fact, some DNN researchers have questioned the efficiency of spectral features derived from critical band theory. In [36], error rates were shown to improve by using a filterbank learning approach rather than just having a fixed set of filters. Therefore, [36] was the first computational study to contradict the energy detector or power spectrum models of critical band theory using purely quantitative and statistical results.

In the present study, we compared masking in cochlear implants and □□ (FBANK+TBANK) input representations in deep learning. The data provides statistical evidence supporting the efficiency of profile analysis, temporal critical bands, and transition bandwidths. Therefore, results in both human hearing and machine hearing oppose the historically accepted critical band theory. Furthermore, all of these findings decrease confidence in previous estimates of peripheral filtering as presumed [3, 19, 22] and adopted in [27, 28]. Moreover, the similarity and compatibility of the results in both human and machine hearing could provide new insight into the ability to process sound and may lead to advances in cochlear implant methods [44] or alternative neural network architectures [45]. For example, [36] indicated that using a nonlinear perceptually motivated log function was appropriate in deep learning, since their results showed that using log nonlinearity with positive weights was preferable.

4. Conclusions

In this chapter, we presented psychoacoustic results that support a recent theory in auditory processing [7, 23]: that the auditory system is actually composed of multiple filterbanks in the processing of sound (instead of just a solitary peripheral filterbank as previously assumed). Psychoacoustic results using electric stimulation in cochlear implant users suggest distorted *profile analysis* (where users are unable to adequately use across-channel cues to compare the shape of the output of different auditory filters), a reliance on *temporal critical bands* (where users relied on filter bandwidths that were consistently broader than predicted by critical band theory), rapidly changing *transition bandwidths* (where users employed separate and different auditory processes to handle either narrow or wide bandwidths), and a distorted *volley theory* of encoding (where users were unable to combine phases of action potentials to analyze a greater frequency of sound). In addition, the results from our deep learning system confirmed the computational effectiveness of combining both spectral filterbanks (FBANK) and temporal filterbanks (TBANK). The combined input representations (each with its own filtering properties) are formed into □□ (double-BANK) features to improve the processing of information in multiple parallel processes. These □□ features all outperformed FBANK features in deep neural network (DNN) systems.

Glossary

Human hearing research

Auditory masking: perceptual phenomenon that occurs when the threshold of audibility for one sound is raised in the presence of another sound.

Cochlear implant: a surgically implanted electronic device that restores a sense of hearing.

Critical band theory: estimates the bandwidth of spectral frequencies within which a second sound is predicted to interfere with the perception of the first sound by auditory masking.

Electric-on-acoustic masking: reproduction of masking by using cochlear implant electrodes.

Electric-on-electric masking: production of masking using only cochlear implant electrodes.

Energy detector model: the nonlinear power spectrum model approximation of auditory responses (which is the inspiration for acoustic features such as FBANK and MFCC).

Place theory: pitch perception depends on the location along the basilar membrane.

Profile analysis: a signal is detected by noting a change in the spectrum at some frequency.

Temporal critical bands: critical bandwidth for a temporal process (e.g., temporal envelope).

Temporal theory: pitch perception depends on the temporal firing patterns of neurons.

Transition bandwidths: occurrence of an interplay between spectral and temporal processes.

Volley theory: groups of neurons respond to a sound by firing action potentials slightly out-of-phase to encode a greater representation of sound that is sent to the brain.

Machine hearing research

Automatic speech recognition: a computational method that allows recognition of language.

Data compression: algorithm that reduces the audio transmission and storage requirements.

Deep learning: branch of machine learning that models high level abstractions in the data via multiple layers of processing and nonlinear transformations within hierarchal structures.

Deep neural networks: artificial neural network inspired by the hierarchal modeling of brains.

Double-BANK (□□) features: the combination of features (e.g.: FBANK+TBANK) as inspired by the observance of temporal critical bands and transition bandwidths in the auditory system.

Frequency amplitude modulation encoding (FAME): alternative features derived via TBANK.

Gaussian mixture model hidden Markov model (GMM-HMM): Bayesian method to align DNNs.

Spectral filterbanks: acoustic features (MFCC or FBANK) inspired by the nonlinear spacing of power spectrum or energy detector models, and critical band theory.

Temporal filterbanks: acoustic features (e.g.: TBANK) inspired by temporal critical bands.

Acknowledgements

In this chapter, some parts of this work was performed when the author was with the Hearing and Speech Laboratory, University of California, Irvine, USA, with Fan-Gang Zeng. The author is now working with the Center for Information Technology Innovation (CITI), Academia Sinica, Taipei, Taiwan, with Yu Tsao.

Author details

Payton Lin

Address all correspondence to: paytonlin20@gmail.com

Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

References

[1] Wegel, R. L., & Lane, C. E. (1924). The auditory masking of one pure tone by another and its probable relation to the dynamics of the inner ear, *Phys. Rev*., 23, 266–285.

[2] Zwislocki, J. J. (1972). A theory of central auditory masking and its partial validation, *J. Acoust. Soc. Am*., 52 (2B), 644–659.

[3] Fletcher, H. (1940). Auditory patterns, *Rev. Mod. Phys*., 12, 47–65.

[4] Berg, B. G. (2004). A temporal model of level-invariant, tone-in-noise detection, *Psychol. Rev*., 111, 917

[5] Spiegel, M. F., & Green, D. M. (1982). Signal and masker uncertainty with noise maskers of varying duration, bandwidth, and center frequency, *J. Acoust. Soc. Am.*, 71(5), 1204–1210.

[6] Shim, A. I., & Berg, B. G. (2013). Estimating critical bandwidths of temporal sensitivity to low-frequency amplitude modulation, *J. Acoust. Soc. Am*., 133(5), 2834–2838.

[7] Berg, B. G. (2007). Estimating the transition bandwidth between two auditory processes: Evidence for broadband auditory filters, *J. Acoust. Soc. Am*., 121(6), 3639–3645.

[8] Wever, E. G., & Bray, C. W. (1937). The perception of low tones and the resonance-volley theory, *J. Psychol.*, 3(1), 101–114.

[9] Schroeder, M. R., Atal, B. S., & Hall, J. L. (1979). Optimizing digital speech coders by exploiting masking properties of the human ear, *J. Acoust. Soc. Am*., 66, 1647–1652.

[10] Choi, C. T., & Lee, Y. H. (2012). A review of stimulating strategies for cochlear implants, in Cochlear Implant Research Updates, Dr. Cila Umat (Ed.), InTech Open Access Publisher.

[11] Noble, J. H., Hedley-Williams, A. J., Sunderhaus, L., Dawant, B. M., Labadie, R. F., Camarata, S. M., & Gifford, R. H. (2016). Initial results with image-guided cochlear implant programming in children, *Otol. Neurotol.*, 37(2), e63–e69.

[12] Wilson, B. S. (2013). Toward better representations of sound with cochlear implants, *Nat. Med.*, 19(10), 1245–1248.

[13] Lin, P., Wang, S.-S., & Tsao, Y. (2015). Temporal information in tone recognition, in *IEEE ICCE*.

[14] Lim, H. H., Tong, Y. C., & Clark, G. M. (1989). Forward masking patterns produced by intracochlear electrical stimulation of one and two electrode pairs in the human cochlea, *J. Acoust. Soc. Am.*, 86, 971–980.

[15] Chatterjee, M., & Shannon, R. V. (1998). Forward masked excitation patterns in multi-electrode electrical stimulation, *J. Acoust. Soc. Am.*, 105, 2565–2572.

[16] Lin, P., Turner, C. W., Gantz, B. J., Djalilian, H. R., & Zeng, F.-G. (2011). Ipsilateral masking between acoustic and electric stimulations, *J. Acoust. Soc. Am.*, 130(2), 858–865.

[17] Lin, P., Lu, T., & Zeng, F.-G. (2013). Central masking with bilateral cochlear implants, *J. Acoust. Soc. Am.*, 133(2), 962–969.

[18] Aronoff, J. M., Padilla, M., Fu, Q. J., & Landsberger, D. M. (2015). Contralateral masking in bilateral cochlear implant patients: a model of medial olivocochlear function loss, *PloS one*, 10(3), e0121591.

[19] van Hoesel, R., & Clark, G. (1997). Psychophysical studies with two binaural cochlear implant subjects, *J. Acoust. Soc. Am.*, 102, 495–507.

[20] James, C., Blamey, P., Shallop, J. K., Incerti, P. V., & Nicholas, A. M. (2001). Contralateral masking in cochlear implant users with residual hearing in the non-implanted ear, *Audiol. Neurootol.*, 6, 87–97.

[21] Buss, E., Hall, III, J. W., & Grose, J. H (2013). Monaural envelope correlation perception for bands narrower or wider than a critical band, *J. Acoust. Soc. Am.*, 133, 405–416.

[22] Patterson, R. D., & Moore, B. C. J. (1986). Auditory filters and excitation patterns as representations of frequency resolution, in *Frequency Selectivity in Hearing*, London: Academic Press, pp. 123–177.

[23] Berg, B. G. (2013). A decision weight analysis of transition bandwidths, *J. Acoust. Soc. Am.*, 121(6), 3639–3645.

[24] Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs, *J. Acoust. Soc. Am.*, 102(4), 2403–2411.

[25] Whitmal, N.A., Poissant, S.F., Freyman, R.L., & Helfer, K.S. (2007). Speech intelligibility in cochlear implant simulations: effects of carrier type, interfering noise, and subject experience, *J. Acoust. Soc. Am.*, 122(4), 2376–2388.

[26] Jordan, M.I., & Mitchell, T.M. (2015). Machine learning: trends, perspectives, and prospects, *Science*, 349(6245), 255–260.

[27] Morgan, N., Bourlard, H., & Hermansky, H. (2004). Automatic speech recognition: an auditory perspective, in *Speech Processing in the Auditory system*, S. Greenberg, W.A. Ainsworth, A.N. Popper, & R.R. Fay (Ed.), New York: Springer, 309–338.

[28] Davis, S. B., & Mermelstein, P. (1990). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, in *IEEE Trans. ASSP.*, 28(4), 357–366.

[29] Schädler, M. R., Meyer, B. T., & Kollmeier, B. (2012). Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition, *J. Acoust. Soc. Am.*, 131(5), 4134–4151.

[30] Do, C.-T., Pastor, D., & Goalic, A. (2010). On the recognition of cochlear implant-like spectrally reduced speech with MFCC and HMM-based ASR, *IEEE Trans. Audio, Speech Language Process.*, 18(5), 1065–1068.

[31] Do, C.-T. (2012). Acoustic simulations of cochlear implants in human and machine hearing research, research, in *Cochlear Implant Research* Updates, Dr. Cila Umat (Ed.), InTech Open Access Publisher.

[32] Lin, P., Chen, F., Wang, S.-S., Lai, Y.-H., & Tsao, Y. (2014). Automatic speech recognition with primarily temporal envelope information, in *Proc. Interspeech*.

[33] Gershman, S.J., Horvitz, E.J., & Tenenbaum, J. B. (2015). Computational rationality: a converging paradigm for intelligence in brains, minds, and machines, *Science*, 349(6245), 273–278.

[34] Tenebaum, J. B., (1999). Bayesian modeling of human concept learning, in *Advances in Neural Information Processing Systems, 11 (NIPS-99)*, M.S. Kearns, S.A. Solla, & D.A. Cohn (Ed.), MIT Press, 59–65.

[35] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senoir, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups, *IEEE Signal Processing Magazine*, 29(6), 82–97.

[36] Sainath, T. N., Kingsbury, B., Mohamed, A., & Ramabhadran, B. (2013). Learning filter banks within a deep neural framework, in *IEEE ASRU*, 297–302.

[37] Sainath, T. N., Kingsbury, B., Mohamed, A., Saon, G., & Ramabhadran B. (2014). Improvements to filterbank and delta learning within a deep neural network framework, in *Proc. ICASSP*, 6839–6843.

[38] Lin, P., Lyu, D.-C., Chang, Y.-F., & Tsao, Y. (2015). Speech recognition with temporal neural networks, in *Proc. Interspeech*.

[39] Lin, P., Lyu, D.-C., Chang, Y.-F., & Tsao, Y. (2015). Temporal alignment for deep neural networks, in *Proc. IEEE GlobalSip*.

[40] Sagayama, S., & Itakura, F. (1978). On individuality in a dynamic measure of speech, in *Proc. Spring Meeting of Acoust. Soc. Japan* (in Japanese), 589–590.

[41] Furui, S. (1986). On the role of spectral transition for speech perception, *J. Acoust. Soc. Am.*, 80(4), 1016–1025.

[42] Zeng, F.-G., Nie, K., Stickney, G., Kong, Y.-Y., Vongphoe, M., Bhargave, A., Wei, C., & Cao, K. (2005). Speech recognition with amplitude and frequency modulations, in *Proc. Nat. Acad. Sci. USA (PNAS)*, 2293–2298.

[43] Zeng, F.-G., & Shannon, R.V. (1994). Loudness-coding mechanisms inferred from electric stimulation of the human auditory system, *Science*, 264, 564–566.

[44] Francart, T., & McDermott, H. J. (2013). Psychophysics, fitting, and signal processing for combined hearing aid and cochlear implant stimulation, *Ear Hearing*, 34(6), 685–700.

[45] Sainath, T. N., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A. R., Dahl, G., & Ramabhadran, B. (2015). Deep convolutional neural networks for large-scale speech tasks, *Neural Networks*, 64, 39–48.