INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Identifying Hearing Loss from Learned Speech Kernels

Shamima Najnin[2], Bonny Banerjee[1,2], Lisa Lucks Mendel[3], Masoumeh Heidari Kapourchali[1,2], Jayanta Kumar Dutta[2], Sungmin Lee[3], Chhayakanta Patro[3], Monique Pousson[3]

[1]Institute for Intelligent Systems,
[2]Department of Electrical & Computer Engineering,
[3]School of Communication Sciences and Disorders
The University of Memphis, Memphis, TN 38152, USA
{snajnin, bbnerjee, llmendel}@memphis.edu

Abstract

Does a hearing-impaired individual's speech reflect his hearing loss? To investigate this question, we recorded at least four hours of speech data from each of 29 adult individuals, both male and female, belonging to four classes: 3 normal, and 26 severely-to-profoundly hearing impaired with high, medium or low speech intelligibility. Acoustic kernels were learned for each individual by capturing the distribution of his speech data points represented as 20 ms duration windows. These kernels were evaluated using a set of neurophysiological metrics, namely, distribution of characteristic frequencies, equal loudness contour, bandwidth and Q_{10} value of tuning curve. It turns out that, for our cohort, a feature vector can be constructed out of four properties of these metrics that would accurately classify hearing-impaired individuals with low intelligible speech from normal ones using a linear classifier. However, the overlap in the feature space between normal and hearing-impaired individuals increases as the speech becomes more intelligible. We conclude that a hearing-impaired individual's speech does reflect his hearing loss provided his loss of hearing has considerably affected the intelligibility of his speech.

Index Terms: Acoustic feature learning, spherical clustering, tuning curve, bandwidth, equal loudness contour, audiogram

1. Introduction

In the current state-of-the-art, personalized tuning of hearing devices, such as cochlear implants (CIs) and hearing aids, to optimize the hearing sensations received is a challenging and time-consuming task, even for highly trained and experienced audiologists. Consequently, the benefits of such devices, particularly CIs, are almost never fully utilized. Limited data is the bottleneck; an audiologist can test a patient for only a few parameter combinations in each visit which, even if done judiciously as in [1], is still inadequate to estimate the optimal combination. Since speech, unlike perception, is easily accessible, it can be mined using machine learning algorithms to infer the nature of hearing loss provided the speech of hearing-impaired individuals reflect that nature. This motivates us to investigate whether a hearing-impaired individual's speech reflects his hearing loss.

Studies have found that the deficiencies in hearing for people with significant hearing loss are reflected in their speech [2, 3]. Hornsby et al. [4] found that as the degree of hearing loss increased, speech perception ability in those frequency regions decreased which affected the individual's ability to produce those sounds. Teoh et al. [5] examined physiological, anatomical, and cognitive evidence in prelingually deafened adults and concluded that inadequate auditory input during the early years of speech and language development constituted the primary limiting factor in the intelligibility of speech. The speech production characteristics of individuals with hearing impairment have been described in depth by a number of researchers [2, 6, 7, 8], indicating several notable features that are distinct to this population, including omission, substitution, and place of articulation errors. The frequency of errors increases with the degree of hearing loss. Abnormal voice characteristics such as harshness, breathiness, and hyper- and hypo-nasality may also be present.

However, there are factors that can improve a prelingually deafened individual's speech intelligibility, such as, the ability to make use of available acoustic cues [2] and presence of auditory input through hearing devices together with auditory-oral/linguistic training [9, 10]. Adults who are postlingually deaf and lose their hearing later in life often suffer little or no deterioration in intelligibility, likely because their residual hearing provides sufficient feedback since their mature speech production systems rely more on motor sensory input rather than auditory information to maintain proper control [11, 12, 13].

In this paper, we report our findings on whether a hearing-impaired individual's speech reflects his hearing loss using a novel line of investigation. As subjects, we considered a cohort of 29 adult individuals, both male and female, consisting of 3 normal and 26 with severe-to-profound hearing loss. Among the hearing impaired, 6, 8 and 12 had high, medium and low speech intelligibility respectively, covering the entire intelligibility spectrum. At least four hours of speech data from reading passages was recorded from each subject. Acoustic kernels were learned for each individual by capturing the distribution of his speech data points represented as 20 ms duration windows. These kernels were represented using 4-dimensional feature vectors constituted of four properties of a set of neurophysiological metrics. We show that, for our cohort, these feature vectors can accurately classify hearing-impaired individuals with low intelligible speech from normal ones using a linear classifier (perceptron). However, the overlap in the feature space between normal and hearing-impaired individuals increases as the speech becomes more intelligible. We conclude that a hearing-impaired individual's speech does reflect his hearing loss provided his loss of hearing has considerably affected the intelligibility of his speech.

2. Models and Methods

This section explicates our algorithm for learning acoustic kernels and the neurophysiological metrics using which properties of these kernels will be analyzed.

http://dx.doi.org/10.21437/Interspeech.2016-1488

2.1. Learning Acoustic Kernels

In machine learning, a number of algorithms have been proposed for learning acoustic kernels from audio data. They largely operate on the time-amplitude or the time-frequency representation of the data. Examples of the former include [14, 15, 16, 17, 18] while those of the latter include [19, 20, 21, 22, 23, 24]. Most algorithms for learning kernels are unsupervised. They learn either by minimizing the reconstruction/prediction error (a.k.a. *generative models*) [16, 15, 22, 17, 18, 19, 21, 23, 24] or by capturing the density of the data [20]. Generative models support a sparse [19, 15, 18, 23, 24] or non-sparse [21, 16, 22, 17] encoding of the data. Supervised algorithms, such as [15], optimize a discriminative objective. The kernels are learned to be either shift-invariant (a.k.a. *convolutional*) [14, 15, 18, 19, 23, 24] or not [16, 22, 17, 25, 26, 20, 21]. Typically, the kernels are evaluated in one of three ways: classification accuracy [15, 17, 19, 20, 21, 23, 24], source separation accuracy [14, 22], and in comparison to neurophysiological findings [18, 16].

In this paper, a soft spherical clustering algorithm is used for learning kernels from time-amplitude representation of speech data. Each window of audio constitutes a data point, the dimension of which depends on the length of the window and the sampling frequency. The algorithm captures the density of the data in an unsupervised manner by maximizing the following objective on convergence: $\ell(\mathcal{X}, \mathcal{W}) = \sum_{i=1}^{k} \sum_{x_j \in \mathcal{N}(i)} (x_j \cdot w_i)$, where $\mathcal{X} = \{x_1, x_2, ...x_n\}$ and $\mathcal{W} = \{w_1, w_2, ...w_k\}$ are the set of d-dimensional data points and kernels (or cluster centers) respectively, and $\mathcal{N}(i)$ is the set of data points in the neighborhood of w_i. Each data point and kernel is normalized to zero mean and unit norm. In case of high-dimensional data, such as audio windows, the direction of a data vector is more important than its magnitude [27] which is captured by the cosine similarity. The algorithm learns non-orthogonal and non-shift-invariant kernels that soft partitions the input space on the surface of a d-dimensional hypersphere of unit radius. Unlike many clustering algorithms, the underlying distribution in spherical clustering is arbitrary [28] which is a very desirable property for our application.

2.2. Neurophysiological Metrics

The learned kernels will be analyzed using the following neurophysiological metrics.

Equal loudness contour (ELC). The human ear does not perceive all sounds equally for the different frequencies or sound intensities as the loudness of a pure tone depends on its frequency. An ELC is a curve that indicates the sound pressure levels resulting in perception of the same loudness at different frequencies across the audible spectrum. The lowest contour represents the minimum audible field (MAF), the absolute threshold of hearing. The shape of each contour differs between normal and hearing-impaired individuals. In particular, the slope of the MAF in the higher frequency region is steeper for hearing-impaired individuals as compared to their normal hearing counterparts [29]. The average slope of the MAF between 4 and 8 KHz is computed as: $Slope_{MAF} = \frac{1}{v-u+1} \sum_{i=u}^{v} \frac{y_{i+1}-y_i}{f_{i+1}-f_i}$, where u is the index of the 4 KHz frequency, v is the index of the 8 KHz frequency, f_i and y_i are the frequency and sound pressure level (SPL) respectively at the i^{th} index. All of our subjects have the minima near 4 KHz in their MAF.

Tuning curve (TC). A frequency TC is used to display the auditory threshold at various frequencies for a single auditory neuron. Each nerve fiber has a *characteristic frequency* (CF) where it responds at threshold. At frequencies below 1000 Hz, TCs are symmetric, and at higher frequencies the curves become increasingly asymmetric and are characterized by a very sensitive, frequency-selective tip and long, broadly-tuned tail. A leading cause of hearing loss is hair cell damage. Damage to outer hair cells results in loss of sensitivity leading to a flattened tip of the TC. Loss of inner hair cells allows the TC to maintain its overall shape but there is a loss of sensitivity. Loss of both inner and outer hair cells result in a major loss of sensitivity as well as a much broader shape to the TC. The distribution of CFs will highlight the frequency regions within the audible range where hair cells are damaged or missing; such regions should be larger or more frequent in individuals with severe-to-profound hearing loss than normal ones. The shape of the TC is captured by its bandwidth and Q_{10} value.

Q_{10} value. The sharpness of a TC is determined by the width of the V-shape of the curve relative to the CF which is commonly expressed in terms of the quality (Q) factor. The Q_{10} is typically used; it refers to the point that is 10 dB below the peak. Formally, $Q_{10} = f_C/BW$ where f_C is the CF and BW is the bandwidth. The half-power points are the usual cutoff values which are used to define a bandwidth. Since it is difficult to determine the half-power points of TCs, the points on the curve that are 10 dB up from the minimum point of the TC are used. The bandwidth of a TC provides important information regarding its frequency selectivity; as bandwidth increases, frequency selectivity decreases. Thus, hearing-impaired individuals ought to have greater bandwidth than their normal counterparts which can be captured by the mean bandwidth of all TCs across the spectrum. For a particular CF, narrower the bandwidth, larger is the Q_{10} dB value. Due to greater bandwidth, the slope of Q_{10} values increases slower with frequency for hearing-impaired individuals as compared to normal-hearing ones.

2.3. Data

All subjects participating in this research had significant hearing loss in both ears with documented speech production errors. All subjects were in reportedly good physical health with no physical, mental, cognitive or emotional limitations. Also, all of them were native speakers of General American Dialect. To be included as a participant in this study, subjects had to have at least a severe sensorineural hearing loss bilaterally with a minimum three-frequency pure tone average of 70 dB HL, have normal middle ear function at the time of testing, and be able to read words and sentences in order to complete the required tasks. In addition, three participants had normal hearing.

Testing was conducted in a double-walled sound-treated booth meeting ANSI Standard S3.1-1999 [30] maximum permissible ambient noise levels for audiometric test rooms. Hearing evaluations were performed using a GSI 61 audiometer and supra-aural TDH-50 headphones, and middle ear function was assessed using a GSI AutoTymp 38 tympanometer. Following the administration of baseline speech production and speech perception tests, subjects read several standardized reading passages which comprised at least four hours of reading material depending on the speed of one's reading. All subjects read exactly the same material. All reading material was recorded in mp3 and wav formats using two separate Marantz digital recorders. Based on subjective evaluation of the speech from

the 26 hearing-impaired subjects by multiple normal-hearing listeners, the subjects were divided into high, medium and low intelligibility categories.

2.4. Perception and Production Measurements

Speech production measurement. Speech intelligibility refers to the proportion of a speaker's output that a listener can readily understand. To estimate the speech intelligibility of the hearing-impaired subjects, the speech of a normal subject is taken as the reference. Each sentence from the reference and the hearing-impaired speech are aligned using dynamic time warping. Speech intelligibility, measured as normalized subband envelope correlation (nSec), is calculated as in [31].

Hearing measurement. Each subject's hearing was quantified by calculating the pure tone average (PTA) which provides the average of the hearing threshold levels at 500, 1000, and 2000 Hz. This frequency region is commonly referred to as the speech frequency region of the audiogram. The PTA is a decibel level that quantifies the degree of hearing loss for each ear.

Perception measurement. All subjects' speech perception ability was evaluated using the AzBio sentences [32]. The AzBio sentences are recorded by both male and female talkers and are routinely used to evaluate the speech perception capabilities of hearing-impaired subjects. All subjects listened to three 20-sentence AzBio lists, one in quiet and two in noise, and listeners were required to repeat the sentences heard. Listener responses were scored as percent correct based on the number of words repeated correctly across all sentences in a list.

3. Experimental Results

All recorded data was downsampled from 44.1 to 16 KHz. The kernels are learned from normalized time-amplitude speech windows of 20 ms duration with 10 ms overlap between consecutive windows. Hence the learned kernels are also time-amplitude signals; they resemble the gammatone filters. The frequency components of a kernel determine its tuning properties, with the most dominant component being its CF. In order to characterize the kernels learned from each of our subjects, we construct a 4-dimensional feature vector $\langle Slope_{MAF}, LossCF, Slope_{Q_{10}}, BW_{avg} \rangle$ where $LossCF$ is the sum of frequency intervals where CFs are absent, $Slope_{Q_{10}}$ is the slope of the linear regression from the Q_{10} vs. CF plot, and BW_{avg} is the mean of the bandwidths of all TCs. As discussed in Section 2.2, these four are identified from the literature as salient features that clearly discriminate between normal and hearing-impaired individuals based on their tuning properties in the peripheral auditory pathway. When trained with this feature vector, a perceptron was successful in finding a linear classification boundary between the normal and low intelligibility subjects in our cohort. However, that was not the case between normal and medium or high intelligibility subjects.

The mean of each of the four features for the normal and hearing impaired with low intelligibility subjects are separated significantly (see Table 1). However, the separation is not so clear between the normal and hearing impaired with high intelligibility subjects, which is expected as the features are derived from their speech and not hearing. Also, consistent with previous findings, the mean of $Slope_{MAF}$, $LossCF$ and BW_{avg} increases while that of $Slope_{Q_{10}}$ decreases as we move from normal to hearing impaired with low intelligibility subjects.

Figure 1 shows the location of each subject in the plots for perception (AzBio, PTA) and speech intelligibility (nSec) scores with respect to the four features. Note that in these plots, the normal subjects are frequently clustered together with minimal variability. Our hearing impaired subjects have higher $Slope_{MAF}$ than normal ones due to higher threshold in the high frequency region, which is consistent with the findings in [29]. Hearing impaired subjects are expected to have more damaged hair cells which is represented by intermittent lack of coverage along the frequency range. The nSec vs. $LossCF$ plot reveals that even though some of the hearing impaired subjects have high speech intelligibility, they produce less number of frequencies in their speech leading to higher $LossCF$ than normal subjects. Soft spherical clustering was able to capture these individual hearing loss characteristics from their produced speech.

Table 1: Mean and standard deviation (μ, σ) of the four features for our cohort of 29 subjects.

	Normal Hearing & Speech (3)	Hearing Impaired (26)		
		High Int. (6)	Medium Int. (8)	Low Int. (12)
$Slope_{MAF}$	0.08, 0.01	0.39, 0.48	0.26, 0.26	0.53, 0.26
$LossCF$	616, 125.83	1616, 964.19	1981, 1184	3012, 1184
$Slope_{Q_{10}}$	6.9, 3.66	8.87, 5.19	6.74, 6.33	0.59, 6.34
BW_{avg}	741.64, 357.74	633.73, 215.25	939.6, 649.39	2169, 649.38

The threshold between normal and hearing impaired subjects, as depicted by the vertical dotted line in each plot in Figure 1, shows that the normal hearing and hearing impaired with low intelligibility subjects are almost separable. Note that subjects 3, 19 and 21 belong to the wrong side in a number of plots. In subjects 3 and 21, very few learned kernels have a characteristic frequency above 1 KHz. Thus, their BW_{avg} and $Slope_{Q_{10}}$ mostly reflect the properties of TCs in the low frequency region which is similar to those of the normal subjects. For subject 19, the bandwidth of the learned kernels in the high frequency region (between 2-6 KHz) turns out to be lower than those in the low frequency region which makes his BW_{avg} similar to normal subjects. Unlike subject 19, the bandwidth increases with characteristics frequency in the normal hearing population.

4. Conclusions

Acoustic kernels were learned using soft spherical clustering from at least four hours of speech data recorded from a cohort of 3 normal and 26 severely-to-profoundly hearing-impaired subjects with different degrees of speech intelligibility. Four neurophysiological features in the peripheral auditory pathway were identified from the literature that differentiate the normal and hearing-impaired individuals. When the learned kernels were represented using the same features, the normal and hearing-impaired individuals with low intelligibility could easily be discriminated. However, the discrimination was less easy between normal and hearing-impaired individuals with medium or high intelligibility. We conclude that a hearing-impaired individual's speech does reflect his hearing loss provided his loss of hearing has considerably affected the intelligibility of his speech.

5. Acknowledgment

This research was supported by NSF grant IIS-1231620.

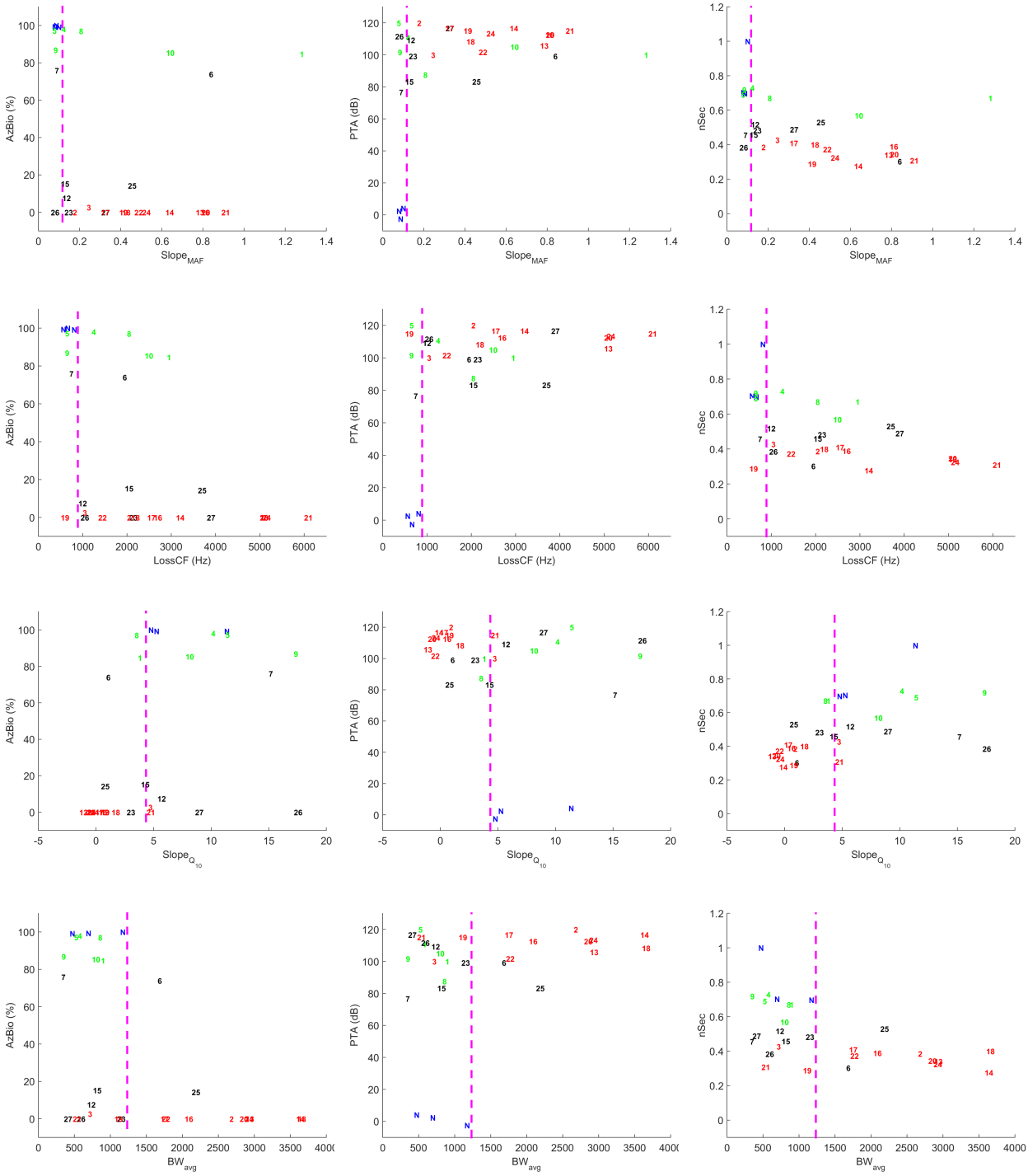

Figure 1: The four features, $Slope_{MAF}$, $LossCF$, $Slope_{Q_{10}}$ and BW_{avg}, are plotted in the four rows (top to bottom) against the three perception and production measurements, AzBio, PTA and nSec, in the three columns (left to right) for our cohort of 29 subjects. Each subject is assigned a unique integer between 1 and 29. Blue integers denote normal subjects while green, black and red ones denote hearing-impaired subjects with high, medium and low speech intelligibility respectively (best viewed in color).The plots of AzBio and PTA vs. $Slope_{MAF}$ (top row) show that normal subjects have lower $Slope_{MAF}$ than hearing-impaired ones with medium or low intelligibility which is consistent with the findings in [29]. $LossCF$ is higher for almost all hearing-impaired subjects than normal ones; the low intelligibility subjects have the highest $LossCF$ followed by the medium intelligibility ones (second row). $Slope_{Q_{10}}$ is lowest for low intelligibility subjects followed closely by some of the medium intelligibility ones while the rest have higher slope (third row). Similarly, BW_{avg} is highest for low intelligibility subjects followed by some of the medium intelligibility ones while the rest have lower average bandwidth (bottom row). The vertical dotted line draws a threshold between normal and hearing impaired subjects.

6. References

[1] B. Banerjee and L. S. Krause, "Automatic performance optimization for perceptual devices," U.S. patent no. 8,401,199, March 2013.

[2] M. J. Osberger and N. S. McGarr, "Speech production characteristics of the hearing impaired," *Speech and Language: Advances in Basic Research and Practice*, vol. 8, pp. 227–288, 1982.

[3] J. Ryalls, A. Larouche, and F. Giroux, "Acoustic comparison of CV syllables in French-speaking children with normal hearing, moderate-to-severe and profound hearing impairment," *J. Multilingual Communication Disorders*, vol. 1, no. 2, pp. 99–114, 2003.

[4] B. W. Hornsby, E. E. Johnson, and E. Picou, "Effects of degree and configuration of hearing loss on the contribution of high-and low-frequency speech information to bilateral speech understanding," *Ear and Hearing*, vol. 32, no. 5, p. 543, 2011.

[5] S. W. Teoh, D. B. Pisoni, and R. T. Miyamoto, "Cochlear implantation in adults with prelingual deafness. part ii. underlying constraints that affect audiological outcomes," *The Laryngoscope*, vol. 114, no. 10, pp. 1714–1719, 2004.

[6] M. Liker, V. Mildner, and B. Šindija, "Acoustic analysis of the speech of children with cochlear implants: A longitudinal study," *Clinical Linguistics & Phonetics*, vol. 21, no. 1, pp. 1–11, 2007.

[7] H. M. Morrison, "The locus equation as an index of coarticulation in syllables produced by speakers with profound hearing loss," *Clinical Linguistics & Phonetics*, vol. 22, no. 9, pp. 726–740, 2008.

[8] M. Hedrick, J. Bahng, D. von Hapsburg, and M. S. Younger, "Weighting of cues for fricative place of articulation perception by children wearing cochlear implants," *Intl. J. Audiology*, vol. 50, no. 8, pp. 540–547, 2011.

[9] S. R. Pratt, "Aural habilitation update: The role of speech production skills of infants and children with hearing loss," *The ASHA Leader*, vol. 10, no. 4, pp. 8–33, 2005.

[10] R. Santarelli, R. De Filippi, E. Genovese, and E. Arslan, "Cochlear implantation outcome in prelingually deafened young adults," *Audiology and Neurotology*, vol. 13, no. 4, pp. 257–265, 2008.

[11] F. H. Guenther, "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production." *Psychological Review*, vol. 102, no. 3, p. 594, 1995.

[12] H. Goehl and D. K. Kaufman, "Do the effects of adventitious deafness include disordered speech?" *J. Speech and Hearing Disorders*, vol. 49, no. 1, pp. 58–64, 1984.

[13] J. Perkell *et al.*, "Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models," *Speech Communication*, vol. 22, no. 2, pp. 227–250, 1997.

[14] T. Blumensath and M. Davies, "Sparse and shift-invariant representations of music," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 50–57, 2006.

[15] R. Grosse, R. Raina, H. Kwong, and A. Y. Ng, "Shift-invariance sparse coding for audio classification," *arXiv preprint arXiv:1206.5241*, 2012.

[16] J.-H. Lee, T.-W. Lee, H.-Y. Jung, and S.-Y. Lee, "On the efficient speech feature extraction based on independent component analysis," *Neural Processing Letters*, vol. 15, no. 3, pp. 235–245, 2002.

[17] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *IEEE Intl. Conf. Acoustics, Speech and Signal Processing*, 2011, pp. 5884–5887.

[18] E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature*, vol. 439, no. 7079, pp. 978–982, 2006.

[19] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in Neural Information Processing Systems*, 2009, pp. 1096–1104.

[20] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *IEEE Intl. Conf. Acoustics, Speech and Signal Processing*, 2015, pp. 171–175.

[21] A. Bertrand, K. Demuynck, V. Stouten, and H. V. Hamme, "Unsupervised learning of auditory filter banks using non-negative matrix factorisation," in *IEEE Intl. Conf. Acoustics, Speech and Signal Processing*, 2008, pp. 4713–4716.

[22] T. Virtanen, "Unsupervised learning methods for source separation in monaural music signals," in *Signal Processing Methods for Music Transcription*. Springer, 2006, pp. 267–296.

[23] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, and Language Processing,*, vol. 20, no. 1, pp. 30–42, 2012.

[24] D. Yu, L. Deng, and G. Dahl, "Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," in *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[25] K. Engan, S. O. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *IEEE Intl. Conf. Acoustics, Speech, and Signal Processing*, vol. 5, 1999, pp. 2443–2446.

[26] W. Dai, T. Xu, and W. Wang, "Dictionary learning and update based on simultaneous codeword optimization (simco)," in *IEEE Intl. Conf. Acoustics, Speech and Signal Processing*, 2012, pp. 2037–2040.

[27] A. Strehl, J. Ghosh, and R. J. Mooney, "Impact of similarity measures on web-page clustering," in *AAAI Workshop on AI for Web Search*, 2000, pp. 58–64.

[28] S. Zhong, "Efficient online spherical k-means clustering," in *Proc. Intl. Joint Conf. Neural Networks*. IEEE, 2005, pp. 3180–3185.

[29] E. Goldstein and J. Brockmole, *Sensation and perception*. Nelson Education, 2016.

[30] ANSI S3. 1-1999 (R2008), "Maximum permissible ambient noise levels for audiometric test rooms," 1999.

[31] J. B. Boldt and D. P. W. Ellis, "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *17th European Signal Processing Conf.* IEEE, 2009, pp. 1849–1853.

[32] A. J. Spahr *et al.*, "Development and validation of the AzBio sentence lists," *Ear and Hearing*, vol. 33, no. 1, p. 112, 2012.