Processing speech signal using auditory-like filterbank provides least uncertainty about articulatory gestures

Prasanta Kumar Ghosh^{a)}

*Signal Analysis and Interpretation Laboratory, Department of Electrical Engineering,
University of Southern California, Los Angeles, California 90089*

Louis M. Goldstein

Department of Linguistics, University of Southern California, Los Angeles, California 90089

Shrikanth S. Narayanan

*Signal Analysis and Interpretation Laboratory, Department of Electrical Engineering,
University of Southern California, Los Angeles, California 90089*

(Received 17 June 2010; revised 19 February 2011; accepted 13 March 2011)

Understanding how the human speech production system is related to the human auditory system has been a perennial subject of inquiry. To investigate the production–perception link, in this paper, a computational analysis has been performed using the articulatory movement data obtained during speech production with concurrently recorded acoustic speech signals from multiple subjects in three different languages: English, Cantonese, and Georgian. The form of articulatory gestures during speech production varies across languages, and this variation is considered to be reflected in the articulatory position and kinematics. The auditory processing of the acoustic speech signal is modeled by a parametric representation of the cochlear filterbank which allows for realizing various candidate filterbank structures by changing the parameter value. Using mathematical communication theory, it is found that the uncertainty about the articulatory gestures in each language is maximally reduced when the acoustic speech signal is represented using the output of a filterbank similar to the empirically established cochlear filterbank in the human auditory system. Possible interpretations of this finding are discussed. © 2011 Acoustical Society of America.

[DOI: 10.1121/1.3573987]

PACS number(s): 43.72.Ar [CYE]

Pages: 4014–4022

I. INTRODUCTION

The speech production system acts as the information transmitter and the auditory system as the receiver in human speech communication. The production system uses choreographed and coordinated movements of several articulators, including the glottis, tongue, jaw, and lips, to produce speech. According to articulatory phonology,¹ speech can be decomposed into basic phonological units called articulatory gestures. The auditory system, on the other hand, decodes the message from the received speech signal; this is performed by cochlear filtering of the sound signal and subsequent transduction of the filtered signal information into electrical impulses that get decoded in the brain.² It has been shown that the frequency–bandwidth dependence in the filterbank model of the cochlea has high coding efficiency for conveying maximal information to the brain for a wide range of natural sounds and, in particular, speech.^{3,4} It is, however, not clear as to what principles have led to the development of the specific articulatory gestures involved in speech production.

Simulations of the self-organization of human-like phonological systems have demonstrated the joint influence of the articulatory and auditory systems,⁵ and it has been

argued that each of these systems is adaptively tuned to the other during real-time speech communication,⁶ including the constraints imposed by the mammalian auditory system on phonetic contrasts.⁷ Evidence from functional magnetic resonance imaging (fMRI) studies has shown that listening to speech activates motor areas involved in speech production.⁸ Several theories of speech perception^{9–11} argue that humans perceive speech by perceiving the articulatory gestures used to produce the speech sound. In spite of much evidence and several theories, how the speech production and perception systems are related to each other is not completely understood. It is not clear what underlying theoretical principle governs the development of articulatory gestures involved in speech production and how it is linked to the signal representation in the auditory system. From a communication theoretic perspective, it can be conjectured that the articulatory gestures in speech production and/or the auditory system may have evolved such that the acoustic properties of the speech are tuned to the cochlear filter characteristics for maximizing the information transfer from the speaker to the listener. To establish a communication between a speaker and a listener, the amount of information in a speaker's message should be preserved in the transmitted speech signal; this presumes that similar information should also be available in the articulatory gesture representations, which are viewed as the basic units of speech.¹ Thus, the articulatory gestures and/or auditory filters might have evolved such that

^{a)}Author to whom correspondence should be addressed. Electronic mail: prasantg@usc.edu.

FIG. 1. (Color online) Articulator data from the X-ray MicroBeam speech production database (Ref. 13). (a) Pellets are placed on eight critical points of the articulators viewed on the midsagittal plane of the speaker, namely, upper lip (UL), lower lip (LL), tongue tip (T1), tongue body (T2 and T3), tongue dorsum (T4), mandibular incisors (MNI), and mandibular molars (MNM). The XY coordinate values of these pellets are recorded. (b) The X and Y coordinate values of eight pellets are shown when a male speaker was uttering “but special.”

the generated speech, after being filtered by the cochlea, can be used in the brain, with the least uncertainty, to decode the speaker’s message or the information preserved in the articulatory gesture representation. This means that the mutual information (MI) between the signal representation at the cochlear output and the articulatory gestures is maximal, in an information theoretic sense.¹² Therefore, to test our hypothesis, we proceed to find the filterbank structure whose output has maximal MI with the articulatory gestures.

We use articulatory movement recordings during speech production to represent articulatory gestures. Since articulatory gestures can vary across languages, for our experiment, we consider articulatory movement data and the concurrently recorded speech signal from three different languages: English, Cantonese, and Georgian. We begin with the description of the datasets used in our experimental analysis.

II. DATASET

We have used the X-ray MicroBeam speech production database collected at the University of Wisconsin¹³ for the experiments related to the English language. This corpus provides temporal trajectories of the movement of the upper lip (UL), lower lip (LL), tongue tip (T1), tongue body (T2 and T3), tongue dorsum (T4), mandibular incisors (MNI), and mandibular molars (MNM) of speakers obtained using the X-ray microbeam technique¹³ during their speech. We have chosen 40 subjects (17 males, 23 females) in this database for our study from 47 available subjects (since the X-ray MicroBeam database suffers from missing sensor positions, we exclude subjects for whom more than 50% of their data are missing). The approximate locations of the articulators are shown in Fig. 1(a) viewed on the midsagittal

plane of a speaker. The locations of each of these eight articulators at any given time are represented by the X and Y coordinates in the midsagittal plane, and the articulator trajectories are sampled at 145.65 Hz. These are obtained using the X-ray microbeam technique¹³ when the subject speaks. For our analysis, we downsampled the microbeam pellet data to a rate of 100 samples/s. A few coordinate values of some pellets at some time points are missing. If any pellet data at any time point are missing, we discard all other pellet data at that time point.

We also include data from two languages distinct from English, namely Cantonese¹⁴ and Georgian,¹⁵ to explore the generalization of the proposed experiments. Unlike English, Cantonese is a tonal language. The consonants and vowels in Cantonese and Georgian are different from those in English, and this implies that the languages employ different articulatory gestures and/or different combinations of gestures. We expect that the kinematics in those languages will reflect the gestural patterns specific to the respective language. The articulatory movements of three (two male and one female) Cantonese and two (both male) Georgian speakers recorded at 500 Hz using the EMMA (Ref. 16) technique, while they spoke, are used as a secondary data source for our study. The recorded articulators are UL, LL, jaw (JAW), tongue tip (TT), tongue body (TB), and tongue dorsum (TD). Similar to the articulatory data from English speakers, we preprocessed the EMMA data from Cantonese and Georgian speakers to achieve a frame rate of 100 Hz.

Note that, in parallel to the articulatory movements in the corpora described above, the speech signal is recorded at 21 739 Hz for English subjects and 20 000 Hz for Cantonese and Georgian subjects. We downsampled each speech recording to 16 kHz for our analysis. Using these parallel

FIG. 2. Parametric representation of filterbanks with different center frequency and bandwidth relationships. (a) The red dashed curve is a warping function which maps the linear frequency axis ω to a warped frequency axis ω_α such that a uniform filterbank on the ω_α axis is equivalent to a nonuniform filterbank on the ω axis, which has a center frequency and bandwidth relationship identical to that of a cochlear filterbank. (b) The warping function is parameterized by the parameter α . Different warping functions are shown for different choices of α (blue curves). The red dashed curve shows the warping function corresponding to the cochlear filterbank and is identical to the red dashed curve in (a).

acoustic and articulatory data from Cantonese and Georgian, we will be able to test our hypothesis in a scenario where the acoustic properties of sounds and the respective articulatory gestures are different from those in English. In Secs. III and IV, we describe how the articulatory corpora are used for specifying the articulatory representations during speech production and how the signal representation at the output of a generic filterbank is specified.

III. ARTICULATORY GESTURE REPRESENTATION

Speech gestures can be modeled as the formation and release of constrictions¹⁷ by particular constricting organs of the vocal tract (lips, TT, etc.). The unfolding of these constrictions over time causes motion in the vocal tract articulators, whose positions are tracked using markers in the X-ray and EMMA data. For example, Fig. 1(b) illustrates the X and Y coordinate trajectories of eight articulators corresponding to an English male speaker's utterance of "but special." We make a generic assumption that the trajectories of UL, LL, T1, T2, T3, T4, MNI, and MNM provide information about critical articulatory gestures involved in producing various speech sounds.¹⁸ For example, in Fig. 1(b), the Y coordinates of UL and LL decrease and increase, respectively, to create the lip closure gesture while producing the sound /b/ (at around 0.2 s) in the word "but" [this is indicated by ↓ and ↑ in Fig. 1(b)]. The tongue tip goes up to the palate and creates constriction for producing /t/ (at around 0.53 s); this gesture is indicated by the peak in the Y coordinate of tongue tip T1 [this is indicated by ↑ in Fig. 1(b)]. We use all the measured movement coordinate values and construct a 16-dimensional vector (Y) as a representation of articulatory gestures every 10 ms.

In a similar fashion, the coordinate values of the available articulators of subjects in the Cantonese and Georgian languages are used to construct an articulatory position vector (Y) as a representation of articulatory gestures.

IV. ACOUSTIC SIGNAL REPRESENTATION

In the auditory system, it is well known that the basilar-membrane in the cochlea performs a running spectral analysis on the incoming sounds.¹⁹ This process can be conceptualized as a bank of tonotopically organized cochlear filters operating on the speech signal. It should be noted that the physiological frequency analysis is different from the standard Fourier decomposition of a signal into its frequency components. A key difference is that the auditory system's frequency response is not linear,¹⁹ e.g., a change from 500 to 1000 Hz is not perceptually equivalent to a change from 5000 to 5500 Hz. The relationship between the center frequency of the analysis filters and their locations along the basilar-membrane is approximately logarithmic in nature.¹⁹ Also, the bandwidth of the filter is a function of its center frequency.²⁰ The higher the center frequency, the wider the bandwidth. Thus on the normal linear frequency axis ω , the filters look as shown in Fig. 2(a).

These depictions allow for a parametric stylization of the cochlear filters where the bandwidths of the brick-shaped filters represent the equivalent rectangular bandwidths (ERB) (Ref. 20) of the cochlear filters at each chosen center frequency along the frequency axis. Interestingly, these filters can be interpreted as uniform filters on a warped frequency axis ω_α with the appropriate warping (Fig. 2). We parameterize the warping function in such a way that by choosing the amount of warping we can construct a filterbank variant with a different relationship between its center frequencies and bandwidths. We use an all-pass filter based parametric approach for warping the frequency axis.²¹ The mathematical form of the warping between ω_α and ω is

$$\omega_\alpha = \frac{2F_s}{\pi} \arctan \left(\frac{(1-\alpha)}{(1+\alpha)} \tan \left(\frac{\omega\pi}{F_s} \right) \right), \quad (1)$$

where F_s is the sampling frequency of the speech signal and α is the sole parameter to control the amount of warping.

Figure 2(b) illustrates different warping functions plotted for different choices of α . For modeling auditory filters along the bark frequency scale, such an all-pass filter based warping of the linear frequency axis is a common approach.²² We find that, for $F_s = 16$ kHz, the filterbank for $\alpha = -0.6$ closely matches the experimentally determined (canonical) cochlear filterbank [the red dashed line in Fig. 2(a)].²³ The advantage of such a parametric warping is that we can now computationally investigate the effects of different filterbank structures on the signal features computed by them.

We compute features from the speech signal using the energies at the output of the filters (mentioned above) over a short duration every 10 ms. Let $\{x(n): 0 \leq n \leq N-1\}$ be the samples of a segment of signal (over a short duration) at the sampling frequency F_s . In our experiments $F_s = 16$ kHz and $N/F_s = 0.02$ s. The energy of the output of the k th filter is given by

$$S_k = \sum_{l=\eta_k}^{\eta_{k+1}} \left| \sum_{n=0}^{N-1} x[n] \exp^{-j(2\pi/N_F)ln} \right|^2, \quad (2)$$

where N_F is the order of the fast Fourier transform (FFT) for computing the spectrum of the signal $x[n]$. $(F_s/N_F)\eta_k$ and $(F_s/N_F)\eta_{k+1}$ are the lower and upper cutoff frequencies of the k th filter. The logarithm of the energies at the output of the filters in multichannel bandpass filtering on the speech signal has been used widely in speech analysis to provide parametric signal representations.^{24–26} We use 20 filters for our experiment, wherein S_k , $k = 1, \dots, 20$, represent the energies of the outputs of 20 filters.

The variation of S_k across filters is computed by the coefficients c_m as follows:

$$c_m = \sum_{k=1}^{20} (\log S_k) \cos \left[m \left(k - \frac{1}{2} \right) \frac{\pi}{20} \right], \quad m = 1, \dots, L. \quad (3)$$

We use c_m , $m = 1, \dots, L$ as features at the output of the filters. In general, $L \leq 20$; however, most of the high order coefficients are very small, in general, and hence the top 13 coefficients are commonly used for machine speech recognition experiments. Hence we also choose $L = 13$ for our experiment. We construct a 13-dimensional feature vector, whose elements are c_m , $m = 1, \dots, 13$.

We construct a vector by using these features and call it a “feature vector.” We computed the feature vectors on the speech signal concurrently recorded in the articulatory corpora in parallel to the articulator tracking. These features capture the variation in the energy of the signal across frequency. The variation in the energy across frequency describes the spectral shapes of various sounds in the speech signal²⁷ and is useful for distinguishing different phonetic categories. This is why we chose features based on the energies at the output of the filters, although there are various other possible features including time- and phase-based features. By choosing different filterbanks (by invoking different warping factors $= \alpha$) we obtain different candidate spectral

shape analyses and the resultant signal features. Among these candidate filterbanks, we are interested in finding out the specific filterbank structure whose features provide the maximal information about articulation. By computing these features every 10 ms, we obtain a sequence of feature vectors as an equivalent representation of the speech signal. Let us denote the feature vector computed using a filterbank obtained with a warping factor α by the variable X_α .

V. MAXIMIZATION OF MI

Vectors Y and X_α , respectively, provide a quantitative description of articulatory gestures in speech production and the signal representation for the corresponding speech obtained by a filterbank parameterized by α . We are interested in quantifying the amount of information that X_α provides regarding Y so that it indicates a measure of certainty with which the brain can decode the speaker’s message or, equivalently, the information in Y from the filterbank output X_α . We use the MI $I(X_\alpha, Y)$ between two random variables for this purpose.¹² MI measures how much information a random variable can provide about another random variable (see Appendix). In our case, we treat X_α and Y as random quantities because the realizations of the acoustic feature, X_α , and the articulatory position, Y , are not identical when a speaker utters the same sound at different times due to differences in a variety of linguistic, contextual, and environmental factors.

In the absence of any closed-form optimization for finding optimum α (i.e., the filterbank corresponding to maximum MI) due to analytical intractability, we randomly pick a hundred α (between -0.95 and 0.95), estimate 20 different MI values at each α (see Appendix for details), and compute the average MI at every chosen α . Since the standard deviation (SD) of the MI estimates is of the order of 10^{-3} (0.1% of the actual MI values), the optimum α obtained by such approach is deemed adequate for drawing our interpretations and conclusions.

We plot the estimated MI across α (different analysis filterbank choices) using the data corresponding to a randomly chosen English male speaker in Fig. 3(a) and a randomly chosen English female speaker in Fig. 3(c). From the plots, it appears that the curves of MI for both speakers are unimodal and skewed toward negative α . Figures 3(b) and 3(d) plot the warping functions and the filterbank structure for α , which yields maximum MI in Figs. 3(a) and 3(c), respectively. It is clear that the filterbank corresponding to the maximum MI is strikingly similar to the cochlear filterbank for both speakers. Also, the MI value obtained by the fixed empirically established cochlear filterbank is within 1% of the maximum possible MI for both speakers.

Similarly, we plot the estimated MI vs α for three Cantonese and two Georgian subjects in Figs. 4(a)–4(e). The filterbanks (α) corresponding to the maximum MI are similar to the cochlear filterbank ($\alpha = -0.6$) for all subjects. This observation is consistent with our experimental findings using parallel acoustic and articulatory data of two randomly chosen subjects in English (Fig. 3).

To examine the filterbanks corresponding to maximum MI across subjects of all languages (a total of 45 subjects—40 English, 3 Cantonese, and 2 Georgian), in Fig. 5(a), we

FIG. 3. The MI between the acoustic feature and the speech articulation for different filterbanks in cases of an English male speaker [(a)–(b)] and an English female speaker [(c)–(d)]. (a) One hundred different filterbanks were generated by randomly selecting α between -0.95 and 0.95 . MI was obtained by averaging 20 different estimates of MI for each choice of α . The SD of MI over these 20 estimates is of the order of 10^{-3} . Maximum MI 0.2637 occurs for $\alpha = -0.5752$. (b) The warping function and the filterbank are plotted for $\alpha = -0.5752$. The red dashed curve shows the warping function corresponding to the cochlear filterbank. MI computed using cochlear filterbank is 0.2630 (c) and (d), (a) and (b) repeated for the female speaker. Maximum MI 0.4576 occurs for $\alpha = -0.5728$. MI computed using cochlear filterbank is 0.4540.

plot the range of α over which the MI is more than 90% of the maximum possible MI for an individual speaker. The blue dot for each speaker in Fig. 5(a) corresponds to the α , at which maximum MI occurs (i.e., the optimum filterbank). We note that the optimum filterbank for extracting maximum information about articulatory gestures from the acoustic speech signal varies from speaker to speaker and need not be always identical to the canonical cochlear filterbank used in our computation. The vocal tract shape and, consequently, the articulation during speech production, represented by the pellet data, are not identical across speakers, in general. Thus, the speaker-specific optimum receiver (filterbank corresponding to maximum MI) is expected to vary considering the variability in the transmitter (articulatory data during speech production). This conjecture is also reflected in our experiment through different optimum filterbanks corresponding to maximum MI for different speakers in our experimental database. The α corresponding to maximum MI [blue dot in 5(a)] appears to have a bi-modal distribution, i.e., some of the optimal α are closer to -0.6 and others are closer to -0.8 . Interestingly we found that optimal α closer to -0.6 mostly correspond to male subjects and those closer

to -0.8 correspond to female subjects. This is shown in Table I. This could be due to the fact that the spectrum of a voiced sound by a female subject is approximately a warped version of the spectrum of the same sound by a male subject because of the relative difference in their vocal tract length. Nevertheless, what is particularly striking is that both the resultant optimal filterbanks and the canonical cochlear filterbank have a similar relationship to the bark frequency scale. In other words, the filterbanks that maximize MI across various speakers in our experiments have wider bandwidths at higher center frequencies compared to those at lower center frequencies, similar to what is seen in empirically established cochlear filterbank processors. This similarity in relative frequency resolution suggests that the coarse frequency resolution at higher frequencies and finer resolution at lower frequencies is essential to extract maximum information regarding articulation from the signal features at the filterbank output.

It is clear that the empirically established cochlear filterbank [red dashed line in Fig. 5(a)] provides a signal representation which has more than 90% of the maximum possible MI with the articulatory gestures across different

FIG. 4. (Color online) The MI between the acoustic feature and the speech articulation for different filterbanks in cases of Cantonese [(a)–(c)] and Georgian [(d)–(e)] subjects. One hundred different filterbanks (i.e., different α on X-axis of each plot) were generated by randomly selecting α between -0.95 and 0.95 . MI was obtained by averaging 20 different estimates of MI for each choice of α . The SD of MI over these 20 estimates is of the order of 10^{-3} . Maximum MIs in (a)–(e) occur for $\alpha = -0.5728$, -0.5196 , -0.6308 , -0.5752 , and -0.5752 , respectively. Note that $\alpha = -0.6$ corresponds to the empirically established cochlear filterbank. This suggests that the empirically established cochlear filterbank is a near optimal filterbank, whose output provides maximum information about the articulatory gestures.

speakers. Note that the canonical cochlear filterbank used in our experiment is based on the analytical formulae of the critical bandwidths²⁰ which best describe the measured critical bandwidths. However, the measured critical bandwidths vary across subjects.²⁰ The articulatory gestures might have evolved such that even small variations from the optimum filterbank can achieve near-maximum MI.

Since the optimum filterbank for obtaining maximum information about articulatory gestures changes across speakers, we would like to quantify how close the cochlear filterbank is to the optimum filterbanks. If the cochlear filterbank would have been the optimum filterbank for each

speaker, then all estimates of MI would be less than the MI obtained by the cochlear filterbank. Since the filterbanks in our analysis are chosen at random, we calculate the percentage of filterbanks which yield less MI compared to the MI obtained by the cochlear filterbank; these are shown in Fig. 5(b) for each speaker. The dashed line corresponds to 92.20%, the average of all percentages; this means, on average, if the empirically determined cochlear filterbank is used for signal representation, it will yield more MI with articulatory gestures compared to any randomly chosen filterbank with a probability of 0.922. This indicates that the cochlear filterbank output is a near optimum signal representation to

FIG. 5. The MI between the acoustic feature and the speech articulation across different speakers. Along the axis labeled “speaker number,” the first 40 points correspond to the English subjects followed by three Cantonese subjects and then followed by two Georgian subjects. (a) Speaker-specific range of α over which the MI is more than 90% of the maximum MI. All ranges are on the negative side indicating that filterbanks with cochlea-like nonuniform frequency resolution achieve near-maximum MI. Blue dots indicate the filterbank corresponding to maximum MI. (b) Percentages of randomly chosen one hundred filterbanks (FBs), which yield less MI than that obtained by the cochlear filterbank. The dashed line corresponds to an average percentage of 92.20%.

TABLE I. Genders of the subjects corresponding to two different ranges of optimal α ($\alpha > -0.7$ and $\alpha < -0.7$).

	#Male	#Female
$\alpha > -0.7$	17	1
$\alpha < -0.7$	0	22

reduce uncertainty in decoding the information about a speaker's articulatory gestures.

VI. DISCUSSION: INTERPRETATIONS OF THE EXPERIMENTAL FINDINGS

Based on the experimental analysis in Sec. V, we observe that, among different filterbank structures for processing a speech signal, the representation using the output of an auditory-like filterbank provides the maximal information about the articulatory gestures involved in producing the speech signal. Based on this experimental finding, one may conclude that the auditory filterbank may have evolved to maximize the information transfer from the speaker to the listener in different languages in the sense that, using the cochlear filterbank, the listener can obtain maximal information about the speaker's articulatory gestures from the received speech signal and therefore decode the speaker's message underlying the acoustic speech signal with the least uncertainty. Such an interpretation assumes that the articulatory gestures and hence the acoustic characteristics of the generated speech act as the sole cause for the development of the auditory system, i.e., the auditory filterbank has evolved to extract maximum information about articulatory gestures from the acoustic speech signal. However, the auditory system is exposed to other natural sounds as well and hence the development of the auditory system is expected to be influenced by all other different sounds, too. Also, it is known that the structure and function of the cochlea and the auditory system are broadly similar across the mammals,^{28,29} whose sound production systems are quite distinct. Therefore, these experiments with articulatory and acoustic data limited to three languages may not be adequate to support such a conclusion.

On the other hand, it has also been shown³ that the cochlear filters are ideal for efficiently encoding information for natural sounds. This suggests that the articulatory gestures during speech production might have evolved such that the acoustic components of the speech signal are modified in a way that maximizes the transfer of information to the brain using the cochlear filters. Thus, one can conjecture that the speech production system has adapted to match the characteristics of the perception system for the transfer of maximal information within the human speech communication system. Adaptation of the speech production system to the auditory system assumes that the characteristics of the auditory system are the cause for the development of the articulatory gestures during speech production. This is a reasonable assumption because the development of the speech production system is targeted toward the auditory system of the listener since it is only the auditory system which processes, analyzes, and facilitates the interpretation of the speech sig-

nal produced by the speaker. However, to clearly establish the adaptation of the articulatory gestures to the auditory filterbank, one needs to compute different articulatory gestures other than those used for speech production and show that gestures corresponding to the speech in different languages are the ones that maximize the MI. But at this point it is not clear how to generate articulatory movements or gestures for the whole set of nonspeech events and link those events to the auditory processing. In that sense it is not possible to conclude anything about the evolution of the articulatory gestures with any definiteness based on the experimental findings in Sec. V.

Finally, one can also argue that there is no particular direction of causality, i.e., the perception system influencing the development of the production system or vice versa. Rather, both systems may have evolved simultaneously. Just like the auditory system is used for processing natural sounds other than speech, the speech production organs are also used for activities other than producing speech, e.g., eating, swallowing, and, importantly, making nonverbal sounds such as laughter and crying, etc. Therefore, there can be many factors that may have influenced the development of both the speech production and speech perception systems. Nevertheless, speech being the primal medium of human communication, maximization of the information transfer from speaker to listener could be a potential factor in the development of both the speech production and perception systems. Further theoretical and experimental investigations are necessary for a better understanding of the relation between the development of the speech production and auditory systems.

VII. CONCLUSIONS

Our experiments with speech articulation and acoustic data from three languages show that the articulatory movements during speech production and the cochlear filterbank characteristics have an optimal relationship in an information theoretic sense. In summary, our experiments demonstrate that the output of the cochlear filters provides maximal information about the articulatory gestures underlying the received speech signal. The Motor Theory and Direct Realism theories of speech perception⁹⁻¹¹ argue that speech is perceived by means of the articulatory gestures used in the speech production. Our results indicate that to obtain maximal information about the articulatory gestures, a cochlea-like filterbank is an ideal choice for processing speech. Thus our results show that speech gestures and the auditory system are well matched to one another and that the filtering properties of the human auditory system maximally preserve information about speech gestures that Motor Theory and Direct Realism require. It is also well known in automatic speech recognition that auditory transformation of the acoustic signal (like mel-scale or bark-scale) improves performance. But there is not really an account of why this should be the case. Our result gives a fundamental, principled account of this finding. Auditory transformations improve performance because they maximize the articulatory information that speakers transmit.

ACKNOWLEDGMENT

This work was supported by funds from NIH, NSF, and ONR.

APPENDIX

MI between two random variables measures the amount of information a random variable can provide about another variable. If the two random variables are mathematically identical, then knowing one of them is equivalent to having the full information about the other one; thus, in such a case, the MI between two random variables attains maximum value. On the other hand, if two random variables are independent, then knowing one does not provide any information about the other and hence the MI is zero.

Let two random variables U and V have a joint probability mass function $p(u, v)$ and marginal probability mass functions $p(u)$ and $p(v)$. The MI between U and V is defined as:

$$I(U; V) = \sum_u \sum_v p(u, v) \log \frac{p(u, v)}{p(u)p(v)}. \quad (\text{A1})$$

But, in our case, we do not have access to the probability density functions of X_α and Y . Hence we consider MI estimation by quantization of the spaces of X_α and Y . This quantization is performed on the data points in both spaces with a finite number of quantization bins. We then estimate the joint distribution of X_α and Y in the newly quantized finite alphabet space using standard maximum likelihood criterion, i.e., frequency counts,³⁰ and finally apply the discrete version of the MI given by Eq. (A1). More precisely, we know that X_α and Y take values in R^{13} and R^{16} spaces respectively. The quantizations of these spaces are denoted by $Q(X_\alpha): R^{13} \rightarrow \mathcal{A}_x$ and $Q(Y): R^{16} \rightarrow \mathcal{A}_y$, where $|\mathcal{A}_x| < \infty$ and $|\mathcal{A}_y| < \infty$. Then the estimate of MI is given by

$$I(Q(X_\alpha), Q(Y)) = \sum_{q_x \in \mathcal{A}_x, q_y \in \mathcal{A}_y} p(Q(X_\alpha) = q_x, Q(Y) = q_y) \times \log \frac{p(Q(X_\alpha) = q_x, Q(Y) = q_y)}{p(Q(X_\alpha) = q_x)p(Q(Y) = q_y)} \quad (\text{A2})$$

It is well known that $I(Q(X_\alpha), Q(Y)) \leq I(X_\alpha, Y)$, because quantization reduces the level of dependency between the random variables. On the other hand, increasing the resolution of $Q(\cdot)$, implies that $I(Q(X_\alpha), Q(Y))$ converges to $I(X_\alpha, Y)$ as the number of bins tends to infinity.³¹ For both spaces, we perform K-means vector quantization with 128 prototypes, i.e., $|\mathcal{A}_x| = |\mathcal{A}_y| = 128$. Increasing the number of prototypes yields similar result.

The speakers in the speech production databases in three languages^{13–15} have different numbers of parallel Y and X_α vectors depending on the duration of their recordings. However to estimate $I(Q(X_\alpha), Q(Y))$, we pick approximately 100 000 parallel vectors for each speaker so that the amount of data used in our analysis is balanced across speakers.

To calculate a realization of $I(Q(X_\alpha), Q(Y))$ for a speaker, we select parallel Y and X_α vectors of the target speaker and quantize (with random initialization) them to $Q(X_\alpha)$ and $Q(Y)$, which are finally used in Eq. (A2). We repeat this process 20 times for the chosen α and report MI averaged over 20 realizations for each speaker. By computing multiple realizations, we capture the inherent variability in the process of quantizing the articulatory and acoustic space.

In the absence of any closed-form optimization for finding optimum α due to analytical intractability, we randomly pick a hundred α (between -0.95 and 0.95) and compute the average MI at every chosen α . Since the SD of the MI estimates is of the order of 10^{-3} (0.1% of the actual MI values), the optimum α obtained by such approach is deemed adequate for drawing our interpretations and conclusions.

- ¹C. P. Browman and L. Goldstein, "Articulatory gestures as phonological units," *Phonology* **6**(2), 201–251 (1989).
- ²S. S. Stevens and F. Warshofsky, *Sound and Hearing*, revised ed. (Time, Inc., New York, 1980), p. 58.
- ³E. C. Smith and M. S. Lewicki, "Efficient auditory coding," *Nature* (London) **439**, 978–982 (2006).
- ⁴M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neurosci.* **5**(4), 356–363 (2002).
- ⁵P. Y. Oudeyer, *Self-Organization in the Evolution of Speech* (Oxford University Press, New York 2006), pp. 32–51.
- ⁶B. Lindblom, "Role of articulation in speech perception: Clues from production," *J. Acoust. Soc. Am.* **99**, 1683–1692 (1996).
- ⁷P. K. Khul, "Discrimination of speech by nonhuman animals: Basic auditory sensitivities conducive to the perception of speech-sound categories," *J. Acoust. Soc. Am.* **70**(2), 340–349 (1981).
- ⁸S. M. Wilson, A. P. Saygin, M. I. Sereno, and M. Iacoboni, "Listening to speech activates motor areas involved in speech production," *Nature Neurosci.* **7**, 701–702 (2004).
- ⁹A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code," *Psychol. Rev.* **74**, 431–461 (1967).
- ¹⁰A. M. Liberman and I. G. Mattingly, "The motor theory of speech revised," *Cognition* **21**, 1–36 (1985).
- ¹¹C. A. Fowler, "An event approach to the study of speech perception from a direct-realist approach," *J. Phonetics* **14**, 3–28 (1986).
- ¹²T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley Interscience, New York, 1991), pp. 12–49.
- ¹³J. R. Westbury, "X-ray microbeam speech production database user's handbook (version 1.0)," <http://www2.uni-jena.de/~xlsiad/uwxrmbdb.html> (Last viewed June 15, 2010) (1994).
- ¹⁴M. Yanagawa, "Articulatory timing in first and second language: A cross-linguistic study," Doctoral dissertation, Yale University, 2006.
- ¹⁵L. Goldstein, I. Chitoran, and E. Selkirk, "Syllable structure as coupled oscillator modes: Evidence from Georgian vs. Tashlhiyt Berber," in *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany (2007), pp. 241–244.
- ¹⁶S. J. Perkell, M. Cohen, M. Svirsky, M. Matthies, I. Garabietta, and M. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *J. Acoust. Soc. Am.* **92**, 3078–3096 (1992).
- ¹⁷E. L. Saltzman and K. G. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychol.* **1**, 333–382 (1989).
- ¹⁸C. P. Browman and L. Goldstein, "Gestural specification using dynamically-defined articulatory structures," *J. Phonetics* **18**, 299–320 (1990).
- ¹⁹K. Johnson, *Acoustic and Auditory Phonetics*, 2nd ed. (Wiley-Blackwell, Malden, MA, 2003), pp. 51–52.
- ²⁰E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Am.* **68**, 1523–1525 (1980).
- ²¹A. Makur and S. K. Mitra, "Warped discrete-fourier transform: Theory and applications," *IEEE Trans. Circuits Syst., I: Fundam. Theory Appl.* **48**, 1086–1093 (2001).

- ²²J. O. Smith and J. S. Abel, "Bark and ERB bilinear transforms," IEEE Trans. Speech Audio Process. **7**, 697–708 (1999).
- ²³P. K. Ghosh and S. S. Narayanan, "Bark frequency transform using an arbitrary order allpass filter," IEEE Signal Process. Lett. **17**(6), 543–546 (2010).
- ²⁴S. W. K. Fu, C. H. Lee, and O. L. Clubb, "A survey on Chinese speech recognition," Commun. COLIPS **6**(1), 1–17 (1996).
- ²⁵G. M. White and R. B. Neely, "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-24**, 183–188 (1976).
- ²⁶P. Cosi, Y. Bengio, and R. D. Mori, "Phonetically-based multi-layered neural networks for vowel classification," Speech Commun. **9**(1), 15–29 (1990).
- ²⁷J. Picone, "Signal modeling techniques in speech recognition," Proc. IEEE **81**, 1215–1247 (1993).
- ²⁸A. Forge and T. Wright, "The molecular architecture of the inner ear," Br. Med. Bull. **63**(1), 5–24 (2002).
- ²⁹K. K. Glendenning and R. B. Masterton, "Comparative morphometry of mammalian central auditory systems: Variation in nuclei and form of the ascending system," Brain Behav. Evol. **51**(2), 54–89 (1998).
- ³⁰R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley-Interscience, New York, 2000), pp. 85–92.
- ³¹G. A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partition of the observation space," IEEE Trans. Inf. Theory **45**, 1315–1321 (1999).