Sādhanā Vol. 36, Part 5, October 2011, pp. 699–712. © Indian Academy of Sciences

Auditory-like filterbank: An optimal speech processor for efficient human speech communication

PRASANTA KUMAR GHOSH[1], LOUIS M GOLDSTEIN[2] and SHRIKANTH S NARAYANAN[1]

[1]Signal Analysis and Interpretation Laboratory, Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA
[2]Department of Linguistics, University of Southern California, Los Angeles, CA 90089, USA
e-mail: prasantg@usc.edu;louisgol@usc.edu;shri@sipi.usc.edu

Abstract. The transmitter and the receiver in a communication system have to be designed optimally with respect to one another to ensure reliable and efficient communication. Following this principle, we derive an optimal filterbank for processing speech signal in the listener's auditory system (receiver), so that maximum information about the talker's (transmitter) message can be obtained from the filterbank output, leading to efficient communication between the talker and the listener. We consider speech data of 45 talkers from three different languages for designing optimal filterbanks separately for each of them. We find that the computationally derived optimal filterbanks are similar to the empirically established auditory (cochlear) filterbank in the human ear. We also find that the output of the empirically established auditory filterbank provides more than 90% of the maximum information about the talker's message provided by the output of the optimal filterbank. Our experimental findings suggest that the auditory filterbank in human ear functions as a near-optimal speech processor for achieving efficient speech communication between humans.

Keywords. Human speech communication; articulatory gestures; auditory filterbank; mutual information.

1. Introduction

Speech is one of the important means of communication among humans. In human speech communication, the talker produces a speech signal, which contains the message of interest. A listener receives and processes the speech signal to retrieve the message of the talker. Hence, in the standard communication systems terminology, the talker plays the role of a transmitter and the listener plays the role of a receiver to establish the interpersonal communication. It is well known that human speech communication is robust to different speaker and environmental conditions and channel variabilities. Such a robust communication system suggests that the human

speech receiver and transmitter are designed to achieve optimal transfer of information from the transmitter to the receiver.

Figure 1 illustrates a communication system view of the human speech exchange. The talker uses his/her speech production system to convert the message of interest into the speech signal, which gets transmitted. The speech is produced by choreographed and coordinated movements of several articulators, including the glottis, tongue, jaw and lips of the talker's speech production system. According to articulatory phonology (Browman & Goldstein 1989), speech can be decomposed into basic phonological units called articulatory gestures. Thus, articulatory gestures could be used as an equivalent representation of the message that the talker (transmitter) intends to transmit to the listener (receiver). The speech signal transmitted by the talker, in general, gets distorted by environmental noises and other channel factors as the acoustic speech wave travels in the air (channel). This distorted speech signal is received by the listener's ear. The ear works as the signal processing unit in the listener's auditory system. As the speech wave passes through various segments of the ear canal, namely outer, middle, and the inner ear, the speech signal is converted to electrical impulses which are carried by the auditory nerve and, finally, get decoded in the brain to infer the talker's message. For an efficient communication between talker and listener, we hypothesize that the components in the listener's auditory system (receiver) should be designed such that maximal information about the talker's message (transmitter) can be obtained from the auditory system output. In other words, the uncertainty about the talker's message should be minimum at the output of the listener's auditory system.

Among the various components in the human speech communication system's receiver, the cochlea (figure 1) is known to be the time–frequency analyser of the speech signal. The structure and characteristics of the cochlea in a typical listener's auditory system is illustrated in figure 2. Figure 2a shows the coiled and uncoiled cochlea. The basilar membrane inside the cochlea performs a running spectral analysis on the incoming speech signal (Johnson 2003, pp. 51–52). 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 (Johnson 2003, pp. 51–52). The relationship between the centre frequency of the analysis filters and their location along the basilar membrane is approximately logarithmic in nature (Johnson 2003, pp. 51–52). Also,

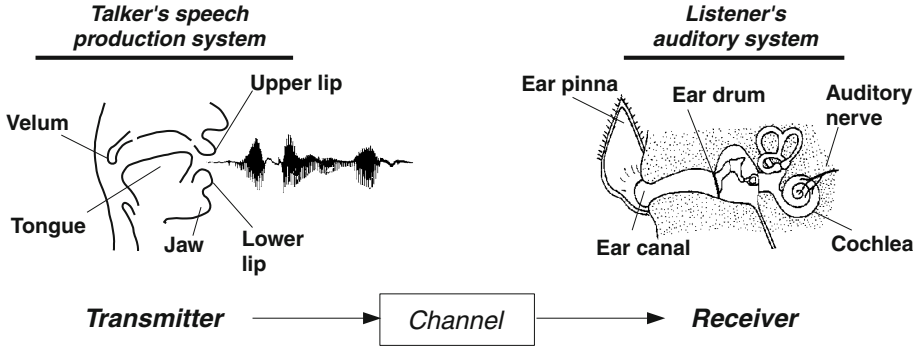

Figure 1. A communication system view of human speech communication. Talker's speech production system consists of speech articulators such as lips, tongue, velum, etc. Outer ear (ear pinna), middle ear and inner ear (cochlea) are the main components of the listener's auditory system (Wikibooks, accessed 13/03/2011).

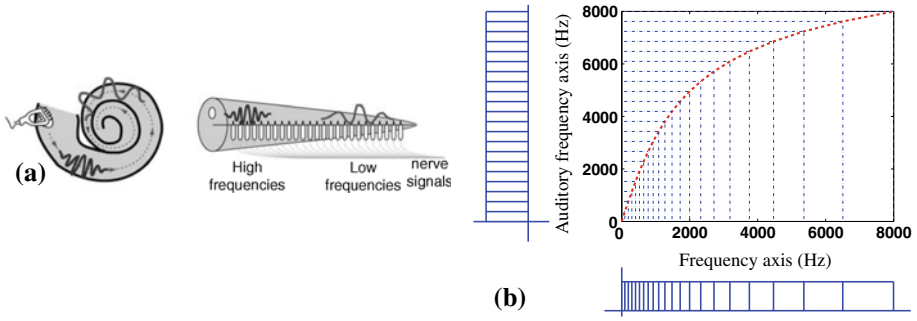

Figure 2. Cochlea and its characteristics. (**a**) The coiled and uncoiled cochlea (Nave, accessed 13/03/2011) with the illustration of the frequency selectivity is shown by damped sinusoid of different frequencies. (**b**) The relation between the natural frequency axis and the frequency map on the basilar membrane (i.e., the auditory frequency axis) from 0 to 8 kHz.

the bandwidth of the filter is a function of its centre frequency (Zwicker & Terhardt 1980). The higher the centre frequency, the wider the bandwidth. A depiction of these filters are shown in figure 2b over a frequency range of 0 to 8 kHz; these ideal rectangular filters are drawn using the centre frequency and the bandwidth data from Zwicker & Terhardt (1980). The bandwidths of the brick-shaped filters represent the equivalent rectangular bandwidths (Zwicker & Terhardt 1980) of the cochlear filters at each chosen centre frequency along the frequency axis. This non-uniform filterbank in the natural frequency axis is referred to as *auditory filterbank*. However, the auditory filterbank appears uniform when plotted in the auditory frequency scale (the frequency scale on the basilar membrane). The relation between the natural and the auditory frequency axis is shown by the red dashed curve in figure 2b. The relationship is approximately logarithmic in nature.

The auditory filterbank is a critical component in the auditory system as it converts the incoming speech signal into different frequency components which are finally converted to electrical impulses by hair cells. The output of the auditory filterbank is used to decode the talker's message. In this study, we focus on the optimality of the filterbank used in the receiver (listener's ear). Our goal is to design a filterbank that achieves efficient speech communication in the sense that the output of the designed filterbank provides the least uncertainty about the talker's message (or equivalently talker's articulatory gestures). Finally, we would like to compare the canonical empirically known auditory filterbank with the optimally designed filterbank to check if the auditory filterbank satisfies the minimum uncertainty principle. We follow the definition of uncertainty from the mathematical theory of communication (Cover & Thomas 1991; Shannon 1948). From an information theoretical viewpoint, it turns out that minimizing uncertainty is identical to maximizing mutual information (MI) between two corresponding random quantities. Therefore, our goal is to design a filterbank such that the filterbank output at the receiver provides maximum MI with the articulatory gestures in the transmitter; we assume that the articulatory gestures can be used as a representation for encoding the talker's message.

For quantitative description of the articulatory gestures, we have to use a database which provides recordings of the movements of the speech articulators involved in speech production (such as tongue, jaw, lips, etc.) when the subject talks. We also need the recording of the speech signal in a concurrent fashion so that the filterbank output can be computed using the recorded speech signal. We describe such a production database in section 2. As articulatory gestures can

vary across languages, in this study, we have used production databases collected from subjects in three different languages – namely, English, Cantonese and Georgian. In sections 3 and 4, we explain the features or the quantitative representations used for the articulatory gestures and the filterbank output. Such quantitative representations are essential for computing MI between them. In section 5, we describe a greedy algorithm for obtaining the optimal filterbank in a language and subject specific manner so that the MI between articulatory gestures and the filterbank output is maximized. We find that the optimal filterbank varies across subjects but the variation is not drastic; rather, the optimal filterbanks are similar to the empirically established auditory filterbank. A communication theoretic explanation for the optimality of the auditory-like filterbank for speech processing is presented in section 6. Conclusions are drawn in section 7.

2. Dataset

We use the X-ray microBeam speech production database collected at the University of Wisconsin (Westbury 1994) for experiments related to the subjects in 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 subjects obtained using X-ray microbeam technique (Westbury 1994) during their speech. We have chosen 40 subjects (17 males, 23 females) from this database for our experiment among available 47 subjects (we exclude a few subjects not having sufficient amount of data required for our experiments). The approximate locations of the articulators are shown in figure 3a viewed on the midsagittal plane of a subject. The locations of each of these eight articulators at any given time are represented by the X and Y co-ordinates in the midsagittal plane, and the articulator trajectories are sampled at 145.65 Hz. For our analysis, we downsampled the microbeam pellet data to a rate of 100 samples/s. A few co-ordinate values of some pellets at some time points are missing. If any pellet data at any time point is missing, we discard all other pellet data at those time points.

We also include data from two languages distinct from English, namely Cantonese (Yanagawa 2006) and Georgian (Goldstein *et al* 2007), to explore the generalization of the proposed experiments. Unlike English, Cantonese is a tonal language. The realizations of phonological units 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. Thus, as a secondary data source for our study, we have used the articulatory movements of three (two male and one female) Cantonese and two (both male) Georgian subjects recorded at 500 Hz using the EMMA (Perkell *et al* 1992) technique, while they spoke. 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 subjects, we pre-processed the EMMA data from Cantonese and Georgian subjects to achieve a frame rate of 100 Hz.

It should be noted that, in parallel to the articulatory movements in the corpora described above, the speech signal is recorded at 21739 Hz for English subjects and 20000 Hz for Cantonese and Georgian subjects. We downsampled the speech signals to 16 kHz for our analysis. Using these parallel acoustic and articulatory data from Cantonese and Georgian, we will be able to examine our communication theoretic hypothesis in a scenario where the acoustic properties of sounds and the respective articulatory gestures are different from those in English. In the following sections, 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.

3. Articulatory gesture representation

Speech gestures can be modelled as the formation and release of constrictions (Saltzman & Munhall 1989) by particular constricting organs of the vocal tract (lips, tongue tip, 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, figure 3b illustrates the X- and Y-co-ordinate trajectories of eight articulators corresponding to a English male subject'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 (Browman & Goldstein 1990). For example, in figure 3b, the Y co-ordinates 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 figure 3b). 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 co-ordinate of tongue tip T1 (this is indicated by ↑ in figure 3b). We use all the measured movement co-ordinate values of eight sensors and construct a 16-dimensional vector (Y) as a representation of articulatory gestures every 10 ms.

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

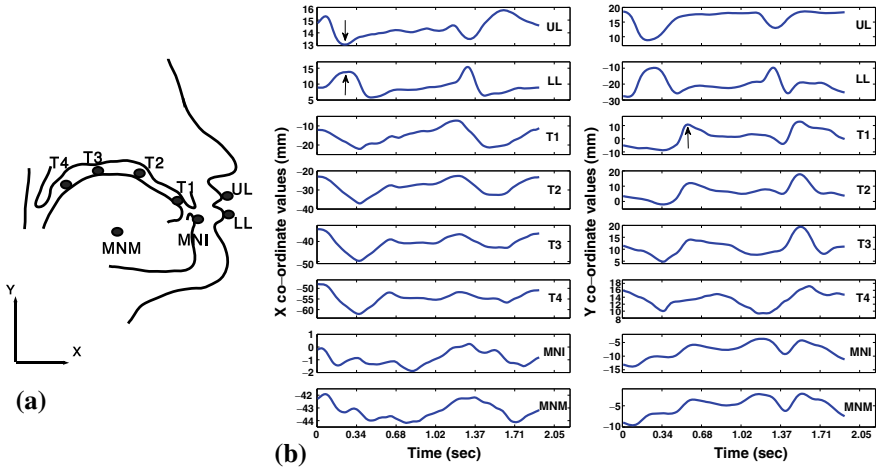

Figure 3. Articulator data from X-ray microBeam speech production database (Westbury 1994). (**a**) Pellets are placed on eight critical points of the articulators viewed on the midsagittal plane of the subject, namely, upper lip (UL), lower lip (LL), tongue tip (T1), tongue body (T2 and T3), tongue dorsum (T4), mandibular incisors (MNI), and mandibular molars (MNM). The XY co-ordinate trajectories of these pellets are recorded. (**b**) The X and Y co-ordinate values of eight pellets are shown when a male subject was uttering 'but special'.

4. Representation of the filterbank output

The filterbank comprising brick-shaped multi-channel bandpass filters can be specified by the upper and lower cut-off frequencies of each bandpass filter. It should be noted that the upper cut-off frequency of a bandpass filter is identical to the lower cut-off frequency of the next bandpass filter in the filterbank (figure 2b). Thus, to specify a filterbank with L filters it is sufficient to know the set of lower cut-off frequencies only; let us denote the set of cut-off frequencies or equivalently the filterbank by B^L.

In the auditory system, the output of the cochlear filterbank gets converted to a series of electrical impulses; however, a complete mechanism for this conversion is not clearly understood. However, it has been reported that the rate of the pulses varies depending on the intensity of the output of the filters (Pathmanathan & Kim 2001; Chatterjee & Zwislocki 1998). Hence, we use the energy of individual filter's output as a representation of the filterbank output. We process the speech signal using each filter and compute the logarithm of the energies at the output of each filter in the filterbank over a short duration every 10 ms. Let $X = \{x(n) : 0 \leq n \leq N-1\}$ be the samples of a segment of the speech signal (over a short duration) at the sampling frequency F_s. In our experiments $F_s = 16$ kHz and $\frac{N}{F_s} = 0.02$ s. The energy of the output of the k^{th} filter is given by:

$$S_k = \log\left(\sum_{l=\eta_{k-1}^L}^{\eta_k^L} \left|\sum_{n=0}^{N-1} x[n]\exp^{-j\frac{2\pi}{N_F}l\eta}\right|^2\right), \quad k = 1, ..., L; \tag{1}$$

where N_F is the order of the Discrete Fourier transform (DFT) for computing the spectrum of the signal $x[n]$. $\frac{F_s}{N_F}\eta_k^L$ and $\frac{F_s}{N_F}\eta_{k+1}^L$ are the lower and upper cut-off frequencies of the k^{th} filter in the filterbank, which comprises L band-pass filters. It should be noted that $B^L = \left\{\eta_0^L,\ \eta_1^L,\ ...\ ,\ \eta_L^L\right\}$, where $\eta_0^L = 0$ and $\eta_L^L = \frac{N_F}{2} - 1$. We construct a vector by using $S_k,\ k = 1, ..., L$ and call it a 'feature vector' X_{B^L} representing the output of the filterbank B^L, i.e., $X_{B^L} = [S_1,\ ...\ , S_L]^{\text{T}}$, where $[\cdot]^{\text{T}}$ is the transpose operator. We compute the feature vectors on the speech signal concurrently recorded with the articulator tracking. By computing these features every 10 ms for a chosen filterbank B^L, we obtain a sequence of feature vectors as an equivalent representation of the filterbank output over time. Depending on the chosen filterbank, the output of the filterbank will vary and hence, the amount of information that this filterbank output can provide about articulatory gestures will also vary.

5. Filterbank optimization

As we have quantitatively defined the description of the talker's articulatory gestures as well as the output of a generic filterbank at the receiver, we can now compute the amount of information that the filterbank output provides about the articulatory gestures. We use the mutual information $I(X_{B^L}; Y)$ between two random variables for this purpose (Cover & Thomas 1991, pp. 12–49). MI measures how much information a random variable can provide about another random variable (see Appendix). In our case, we treat X_{B^L} and Y as random quantities because the realizations of the filterbank output, X_{B^L}, and the articulatory position, Y, are not identical when a subject utters the same sound at different times owing to differences in a variety of linguistic, contextual and environmental factors.

For an efficient communication between the talker and the listener, the filterbank at the receiver (listener) needs to be selected in such a way that its output provides maximum information regarding the talker's articulatory gestures, i.e.,

$$B^{L\star} = \arg\max_{B^L} I(X_{B^L}; Y). \tag{2}$$

It is easy to show that maximizing $I(X_{B^L}; Y)$ is equivalent to minimizing $H(Y|X_{B^L})$, i.e., the conditional uncertainty of articulatory gestures (Y) given the filterbank output X_{B^L} (see Appendix for details).

In our production database (section 2), each subject can be treated as a talker. We optimize the filterbank for each subject in the production database, i.e., we design the filterbank at the receiver for achieving efficient communication with each talker separately in the database. Performing subject-specific optimization provides us the opportunity to analyse the variability in optimal filterbank structure across talkers. For our experiment, we consider speech signals sampled at $F_s = 16$ kHz, i.e., $\frac{F_s}{2} = 8$ kHz. There are 20 critical bands in a range of 0 to 8 kHz as given by the critical bandwidth data (Zwicker & Terhardt 1980). Hence, for the filterbank optimization, we have chosen $L = 20$.

The optimization in Eq. (2) does not have a closed form solution, rather it is a combinatorial problem (and, hence, computationally prohibitive) because the optimization in Eq. (2) is equivalent to finding the cut-off frequencies $\{\eta_k^l\}_{k=0}^L$ from the $\frac{N_F}{2}$ frequency points such that $0 = \eta_0^L < \eta_1^L < \cdots \eta_{L-1}^L < \eta_L^L = \frac{N_F}{2} - 1$. In the absence of any closed-form optimization due to analytical intractability of the objective function (in Eq. (2)), we adopt a greedy algorithm where the optimal filterbank is obtained using a series of decomposition with low-pass and high-pass filters for every frequency band under consideration. This is similar to the dyadic filterbank analysis in wavelets (Strang & Nguyen 1996). A tree-structured illustration of this greedy approach to filterbank optimization is shown in figure 4.

The level 1 in the tree corresponds to the root node. There are $(l-1)$ nodes at the level l ($l > 1$). It should be noted that the root node corresponds to an all-pass filter, i.e., a filter of constant magnitude from 0 to 8 kHz. A node in the level l corresponds to a filterbank having l band-pass filters. These bands at level l are determined by low-pass and high-pass filter decomposition of each band of a filterbank chosen from the previous level (level $l-1$) using MI criterion. There are $(l-1)$ bands in the filterbank chosen from the previous level $(l-1)$; hence, there are $(l-1)$ possible decomposition [or $(l-1)$ possible filterbanks or nodes] at the level l of the tree. Let us denote $(l-1)$ filterbanks in the level l by $B_k^l,\ k = 1, ..., l-1$. For each B_k^l, let $M_k^l \triangleq I(X_{B_k^l}; Y)$. The filterbank (or node) at level $l-1$, which yields the maximum MI among $M_k^{l-1},\ k = 1, ..., l-2$, is used for further decomposition at level l. Nodes corresponding to maximum MI are indicated in figure 4 by circles around the maximum M_k^l in each level. In the filterbank optimization, as our goal is to find the best 20-band filterbank, we run this optimization until the algorithm finds the filterbank with the maximum MI at level 20 in the tree structure. The filterbank with the maximum MI at level 20 is reported as the optimal filterbank. This approach of filterbank optimization is greedy because at every successive level of the tree structure, we select the filterbank which yields the maximum MI. In general, the optimized filterbank can be sub-optimal because of the greedy nature of the optimization. Owing to the lack of any other time-efficient algorithms (apart from full-search) for obtaining optimal solution to the optimization problem (Eq. (2)), we assume that the solution obtained by the proposed greedy algorithm is sufficient for our conclusions and interpretations.

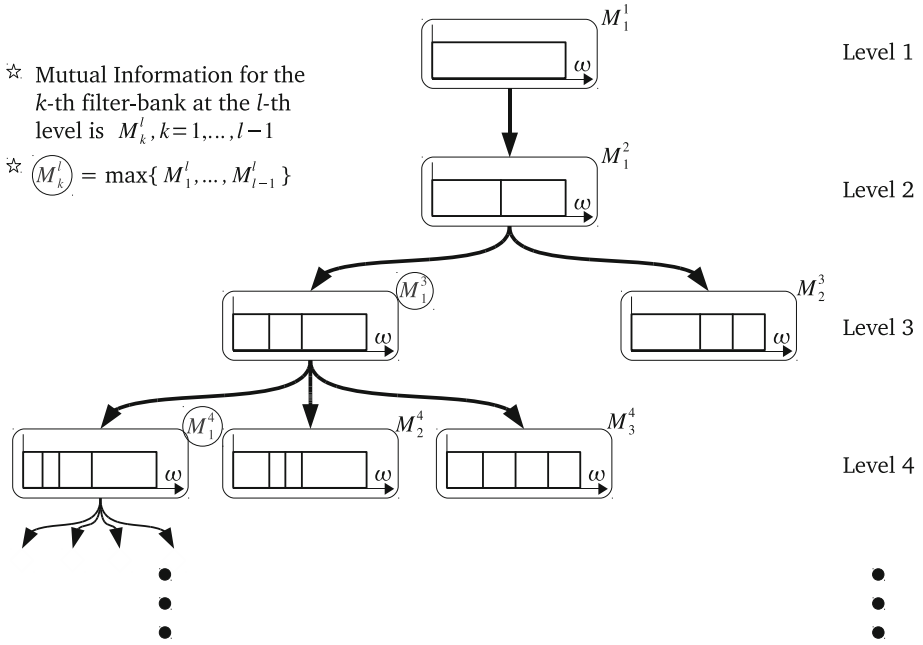

Figure 4. A tree-structure illustration of the greedy approach for optimizing the filterbank.

For each subject in the production database, let us suppose a total of T realizations of articulatory position vectors (section 3) and the corresponding short-time speech segments (20 ms) (section 4) are denoted by Y^j and X^j, $j = 1, ..., T$, respectively. It should be noted that given

Algorithm 1 A tree-structured illustration of the greedy approach to filterbank optimization

$\eta_0^1 = 0, \eta_1^1 = \frac{N_F}{2} - 1, l = 2$

$\eta_0^l = 0, \eta_l^l = \frac{N_F}{2} - 1$ & $\eta_1^l = \frac{\eta_0^{l-1} + \eta_{l-1}^{l-1}}{2}$

$B_1^l = \{\eta_0^l, \cdots, \eta_{l-1}^l\}$; compute $X_{B_1^l}^j$ from X^j, $j = 1, \cdots, T$

Compute $M_1^l = I(X_{B_1^l}; Y)$ using $\left\{\left(X_{B_1^l}^j, Y^j\right);\ j = 1, ..., T\right\}$

for $l = 3$ to 20 **do**

  for $k = 1$ to $l - 1$ **do**

    $B_k^l \leftarrow \left\{B_{k^\star}^{l-1}, \frac{\eta_{k-1}^{l-1} + \eta_k^{l-1}}{2}\right\}$

    Compute $X_{B_k^l}^j$ from X^j, $j = 1, \cdots, T$

    Compute $M_k^l = I(X_{B_k^l}; Y)$ using $\left\{\left(X_{B_k^l}^j, Y^j\right);\ j = 1, ..., T\right\}$

  end for

  $k^\star = \arg\max_k M_k^l$

end for

Return $B_{k^\star}^{20}$ and $M_{k^\star}^{20}$

a filterbank B_k^l we can compute the realization of $X^j_{B_k^l}$ from X^j, $j = 1, ...T$ using the method outlined in section 4. Given $\left\{Y^j,\ X^j,\ j = 1, ..., T\right\}$, the algorithm for filterbank optimization is described in Algorithm 1.

6. Results and discussion

Figure 5a illustrates how the filterbank corresponding to the maximum MI evolves with increasing number of levels in the greedy optimization for a randomly chosen English subject from the production database. The rectangular filterbanks are drawn over the frequency axis (horizontal

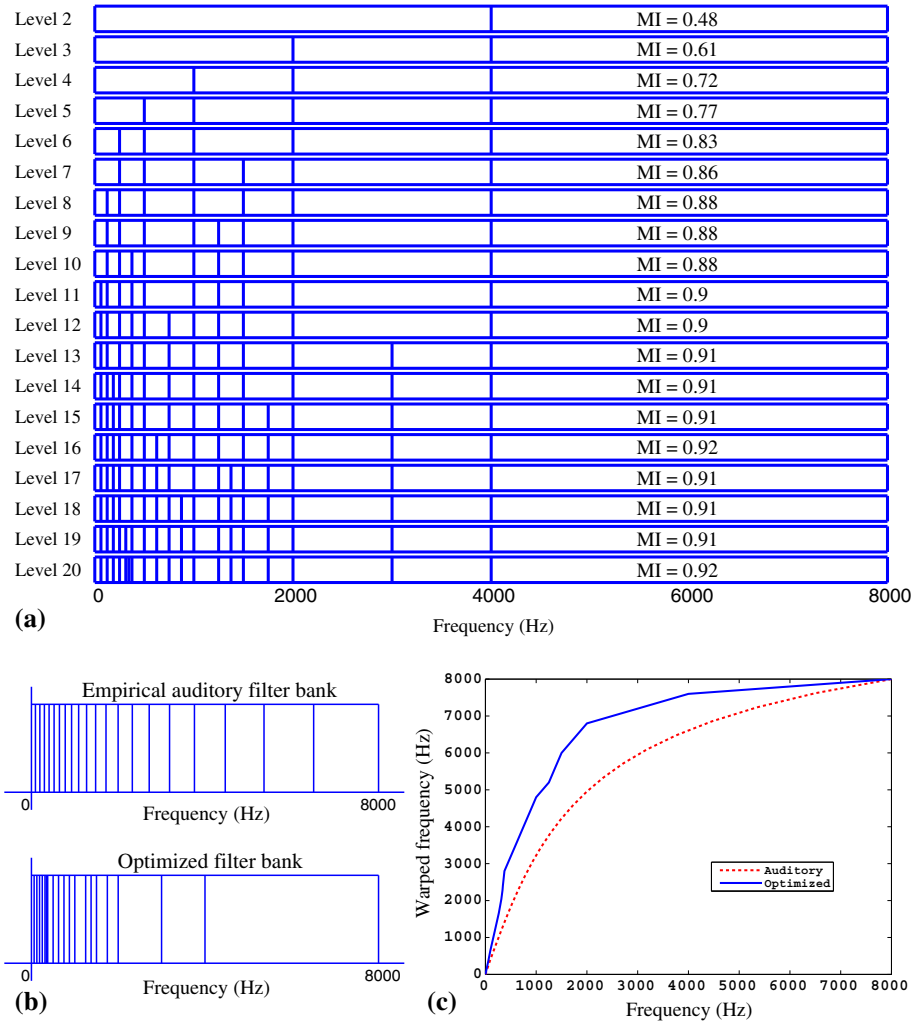

Figure 5. Result of the filterbank optimization (**a**) Filterbank corresponding to maximum MI at each level of the greedy optimization. (**b**) The empirical auditory filterbank and the optimal filterbank (i.e., the filterbank corresponding to the maximum MI at level 20). (**c**) The warping function between the frequency axis and the warped frequency axis obtained by the empirical auditory filterbank and the optimal filterbank.

axis) for increasing levels; the level numbers are indicated on the left side of the drawn filterbanks. The maximum MI corresponding to the filterbank at each level is also shown. It should be noted that the maximum MI at each level need not be strictly increased with increasing level number due to the greedy nature of the algorithm. The algorithm only ensures that the filterbank corresponding to the maximum MI is picked in each level. There is no guarantee that the maximum MI at level l should be greater than that at level $l - 1$. For example, the maximum MI at level 17 is 0.91, whereas that at level 16 is 0.92 in figure 5a. The greedy algorithm can be modified by running it back and forth to ensure strict monotonicity in the MI values across levels; however, it requires much longer optimization time compared to the one presented here.

The optimal filterbank at level 20 is shown in figure 5b below the empirically established auditory filterbank. (The auditory filterbank is identical to the one shown in figure 2b.) It is clear that the filters in the empirical auditory filterbank are not strictly identical to those in the optimal filterbank. However, it is interesting to observe that both the optimal and the empirical auditory filterbank have filters with high frequency resolution at low centre frequencies and low frequency resolution at higher centre frequencies. Thus, the frequency axis corresponding to these filterbanks are warped with respect to the standard frequency axis. This warping function can be obtained using the centre frequencies and the bandwidths of the filters in the filterbank. The warping function for both the auditory and the optimal filterbanks are shown in figure 5c. These warping functions are similar to each other according to the nature of the frequency warping they produce. However, we need to examine the variability of the frequency warping corresponding to the optimal filterbanks across all subjects.

Figure 6 shows the warping functions obtained by using the optimal filterbanks separately for all 45 subjects (40 English + 3 Cantonese + 2 Georgian) against the warping function obtained by the empirically established auditory filterbank (dashed line in figure 6). It is important to

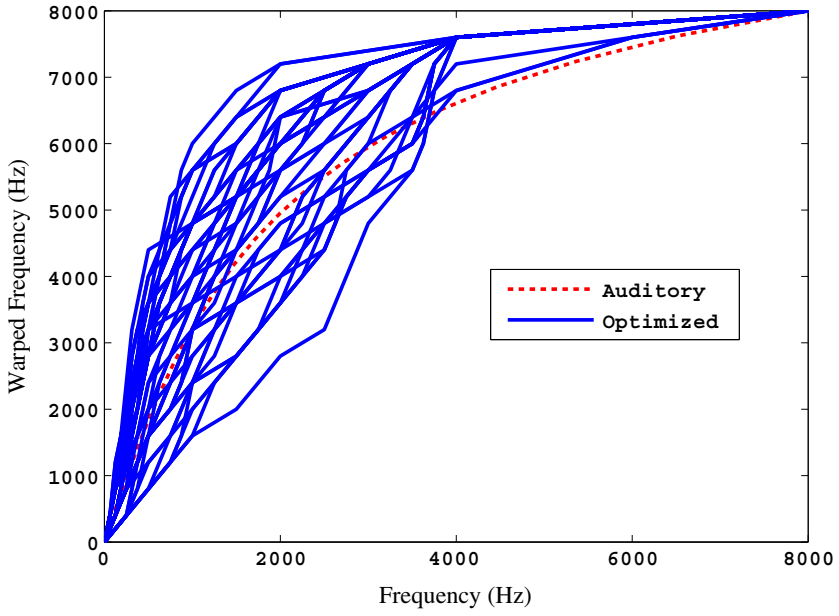

Figure 6. The warping functions corresponding to the optimal filterbanks for all 45 subjects (40 English + 3 Cantonese + 2 Georgian) considered from the production databases.

note that the optimal filterbanks are similar to the auditory one in the sense that the frequency resolution of the filters in the filterbank decreases with the increasing centre frequencies of the filters. This finding is consistent with our earlier studies (Ghosh *et al* 2011), although the filterbank optimization was not performed in a generic fashion rather a specific parameterization of the filterbank was considered. However, the warping functions corresponding to the optimal filterbanks are not identical to the one corresponding to the empirically established auditory filterbank. This can be due to the fact that each optimal filterbank is tuned only for speech signal from a specific subject; however, the auditory filterbank in human ear is exposed not only to speech signal but also to a variety of other sounds including environmental and natural sounds. For example, Smith & Lewicki (2006) demonstrated that the auditory filters are efficient from information encoding perspective considering various natural sounds including speech. As our objective function in this study is derived from a communication theoretic principle for achieving efficient speech communication between a talker and a listener, we have not considered any sounds other than speech. Also the empirically established cochlear filterbank is obtained from the data recorded from different human subjects and there is inherent variability in the critical bandwidths across different human subjects. Thus, the empirically established auditory filterbank just represents a notional filterbank for processing speech and it does not capture the variability in speech processing across subjects. From that perspective, it is expected that the optimal filterbanks corresponding to each subject would not be identical to the empirically established auditory filterbank, rather, on an average, the empirically established auditory filterbank would be similar to the optimal filterbanks.

As the computationally determined optimal filterbanks are not identical to the empirical auditory filterbank, we investigate how close the empirical auditory filterbank is to the optimal

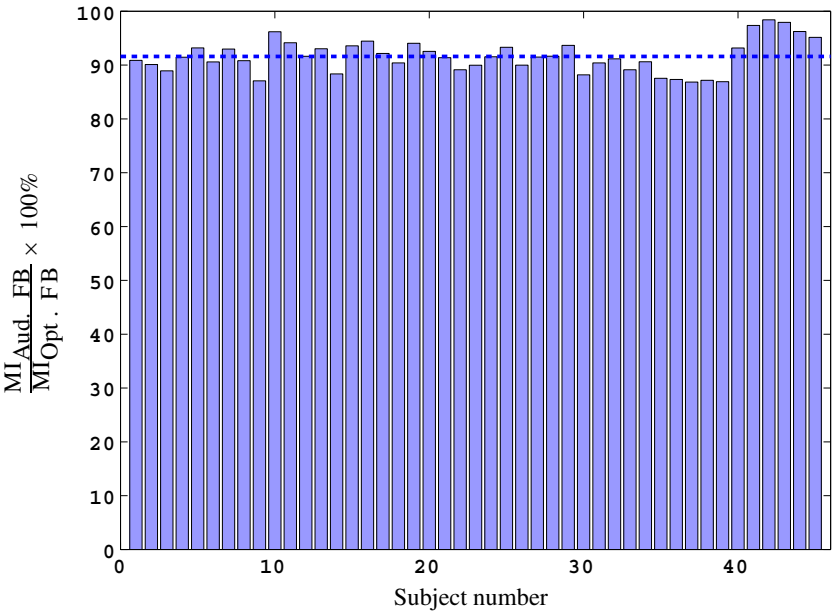

Figure 7. The ratio (in percentage) of the MI computed for auditory filterbank (Aud. FB) and the optimal filterbank (Opt. FB) for all 45 subjects (40 English + 3 Cantonese + 2 Georgian) considered from the production databases. The average of all ratio values (in percentage) is 91.6% (blue dashed line).

filterbank for each subject considered in our experiment; this is done by computing the objective function value $I(X_{B^L}; Y)$ (Eq. (2)) when the empirical auditory filterbank is used as B^L. It should be noted that empirical auditory filterbank is different from the optimal filterbank and, hence, sub-optimal with respect to the objective function in Eq. (2). Let the objective function (or MI) value for empirical auditory filterbank be denoted by M_{Aud}. The optimized filterbank yields the maximum MI at Level 20 of the tree-structured greedy algorithm and the maximum MI value is $M_{k^\star}^{20}$ (see Algorithm 1). We examine the closeness between M_{Aud} and $M_{k^\star}^{20}$ in a subject-specific manner. For each subject in the production database, the ratio $\frac{M_{\text{Aud}}}{M_{k^\star}^{20}}$ is shown by the bar graph in figure 7. It is clear that the ratio $\frac{M_{\text{Aud}}}{M_{k^\star}^{20}} < 1$, i.e., the MI obtained using optimal filterbank is always greater than that obtained using the auditory filterbank. This is expected as the filterbank optimization is performed to maximize the mutual information between filterbank output and the articulatory gestures. It is interesting to observe that although the proposed greedy algorithm may result in sub-optimal solution $B_{L^\star}$, the MI value corresponding to the auditory filterbank never exceeds the MI value corresponding to $B_{L^\star}$ for any subject. Rather M_{Aud} is on an average (across all subjects) 91.6% of the maximum possible MI (i.e., $I(X_{B^{L\star}}; Y)$) achieved by the optimal filterbank. In other words, the output of the auditory filterbank provides more than 90% of the maximum possible information about talker's message (or equivalent articulatory gestures). Thus, the auditory filterbank in the listener's ear acts as a near-optimal speech processor in decoding the talker's message for achieving an efficient speech communication between the talker and the listener.

7. Conclusions

Our experimental result reveals that there is an inherent similarity among the filterbanks optimized in a talker-specific manner for retrieving maximum information about a talker's (transmitter) message from the speech signal. The empirically established auditory filterbank in human ear is similar to the computationally determined optimal filterbanks. This indicates that as far as an efficient speech communication between a talker and a listener is concerned, the auditory-like filterbank offers a near-optimal choice as a speech processing unit at the receiver.

In our experiment, we have used talkers from three different languages (English, Cantonese, and Georgian) with different phonological structure and hence different acoustic characteristics of sounds. The experiment was designed to achieve the best filterbank in the receiver so that maximum information about the transmitter's message can be obtained from the filterbank output. Our results show that, consistently for all languages, the auditory filterbank output provides more than 90% mutual information about the talker's articulatory gestures compared to that provided by a talker-specific optimal filterbank. Thus, from a communication theoretic perspective, the auditory-like filterbank is an optimal choice for processing speech signals and inferring maximal information about the talker's message.

Appendix: Mutual information

Mutual information (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 (Cover & Thomas 1991, pp. 12–49):

$$I(U; V) = \sum_{u} \sum_{v} p(u, v) \log \frac{p(u, v)}{p(u)p(v)}. \tag{3}$$

It is easy to show that $I(U; V) = H(V) - H(V|U)$, where $H(V)$ is the entropy of V and $H(V|U)$ is the conditional entropy of V given U. $H(V|U)$ measures the amount of uncertainty that U provides about V. It should be noted that when $H(V)$ is fixed, maximizing $I(U; V)$ is equivalent to minimizing $H(V|U)$.

For computing $I(X_{B^L}; Y)$ we need to know $p(X_{B^L})$, $p(Y)$, and $p(X_{B^L}, Y)$. However, in our case, we do not have access to the probability density functions of X_{B^L} and Y. Hence, we consider MI estimation by quantization of the spaces of X_{B^L} 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_{B^L} and Y in the newly quantized finite alphabet space using standard maximum likelihood criterion, i.e., frequency counts (Duda & Hart 2000, pp. 85–92) and finally apply the discrete version of the MI given by Eq. (3). More precisely, we know that X_{B^L} and Y take values in $\mathcal{R}^L$ and $\mathcal{R}^{16}$ spaces, respectively. The quantizations of these spaces are denoted by $Q(X_{B^L}) : \mathcal{R}^L \longrightarrow \mathcal{A}_x$ and $Q(Y) : \mathcal{R}^{16} \longrightarrow \mathcal{A}_y$, where $|\mathcal{A}_x| < \infty$ and $|\mathcal{A}_y| < \infty$. Then the estimate of MI is given by:

$$\begin{aligned} I(Q(X_{B^L}); Q(Y)) \;=\; & \sum_{q_x \in \mathcal{A}_x, q_y \in \mathcal{A}_y} p(Q(X_{B^L}) = q_x, Q(Y) = q_y) \\ & \times \log \frac{p(Q(X_{B^L}) = q_x, Q(Y) = q_y)}{p(Q(X_{B^L}) = q_x)p(Q(Y) = q_y)}. \end{aligned} \tag{4}$$

It is well known that $I(Q(X_{B^L}); Q(Y)) \leq I(X_{B^L}; Y)$, because quantization reduces the level of dependency between the random variables. On the other hand, increasing the resolution of $Q(\cdot)$, implies that $I(Q(X_{B^L}); Q(Y))$ converges to $I(X_{B^L}; Y)$ as the number of bins tends to infinity (Darbellay & Vajda 1999). 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 subjects in the speech production databases in three languages (Westbury 1994; Yanagawa 2006; Goldstein *et al* 2007) have different numbers of parallel Y and X_{B^L} vectors depending on the duration of their recordings. However, to estimate $I(Q(X_{B^L}); Q(Y))$, we pick approximately 100,000 parallel vectors of X_{B^L} and Y for each subject so that the amount of data used in our analysis is balanced across subjects.

To calculate a realization of $I(Q(X_{B^L}); Q(Y))$ for a subject, we select parallel Y and X_{B^L} vectors of the target subject and quantize (with random initialization) them to $Q(X_{B^L})$ and $Q(Y)$, which are finally used in Eq. (4). We repeat this process several times for a chosen filterbank B^L to capture the inherent variability in the process of quantizing the articulatory and acoustic space. The standard deviation of the MI estimates is found to be of the order $\sim$0.1% of the actual MI values. Hence, we use the average estimated MI for our experiments.

References

Browman C P, Goldstein L 1989 Articulatory gestures as phonological units, *Phonology* 6(2): 201–251

Browman C P, Goldstein L 1990 Gestural specification using dynamically-defined articulatory structures, *J. Phonetics* 18: 299–320

Chatterjee M, Zwislocki J J 1998 Cochlear mechanisms of frequency and intensity coding. II. Dynamic range and the code for loudness, *Hear. Res.* 124(1–2): 170–181

Cover T M, Thomas J A 1991 *Elements of information theory* (New York: Wiley Interscience)

Darbellay G A, Vajda I 1999 Estimation of the information by an adaptive partition of the observation space, *IEEE Trans. Inform. Theory* 45: 1315–1321

Duda R O, Hart P E 2000 *Pattern classification and scene analysis* (New York: Wiley-Interscience)

Ghosh P K, Goldstein L M, Narayanan S S 2011 Processing speech signal using auditory-like filterbank provides least uncertainty about articulatory gestures, *J. Acoust. Soc. Am.* 129(6): 4014–4022

Goldstein L, Chitoran I, Selkirk E 2007 Syllable structure as coupled oscillator modes: evidence from Georgian vs. Tashlhiyt Berber, *Proc. 16th International Congress of Phonetic Sciences, Saarbrucken, Germany*, pp. 241–244

Johnson K 2003 *Acoustic and auditory phonetics* (MA, USA: Wiley-Blackwell) 2nd edition

Nave C R Place theory. Accessed 13/03/2011. URL http://hyperphysics.phy-astr.gsu.edu/hbase/sound/place.html

Pathmanathan J S, Kim D O 2001 A computational model for the AVCN marginal shell with medial olivocochlear feedback: generation of a wide dynamic range, *Neurocomputing* 38: 807–815

Perkell S J, Cohen M, Svirsky M, Matthies M, Garabieta I, Jackson M 1992 Electro-magnetic midsagittal articulometer systems for transducing speech articulatory movements, *J. Acoust. Soc. Am.* 92: 3078–3096

Saltzman E L, Munhall K G 1989 A dynamical approach to gestural patterning in speech production, *Ecol. Psychol.* 1: 333–382

Shannon C E 1948 A mathematical theory of communication, *Bell Syst. Tech. J.* 27: 379–423

Smith E C, Lewicki M S 2006 Efficient auditory coding, *Nature* 439: 978–982

Strang G, Nguyen T 1996 *Wavelets and filter banks* (Wellesley, MA: Wellesley-Cambridge Press)

Westbury J R 1994 X-ray microbeam speech production database user's handbook version 1.0. http://www2.uni-jena.de/~x1siad/uwxrmbdb.html (date last viewed 6/15/2010)

Wikibooks. Anatomy and physiology of animals/the senses. Accessed 13/03/2011. URL http://en.wikibooks.org/wiki/Anatomy_and_Physiology_of_Animals/The_Senses

Yanagawa M 2006 *Articulatory timing in first and second language: a cross-linguistic study*. Doctoral dissertation, Yale University

Zwicker E, Terhardt E 1980 Analytical expressions for critical-band rate and critical bandwidth as a function of frequency, *J. Acoust. Soc. Am.* 68: 1523–1525