Average Spectrotemporal Structure of Continuous Speech Matches with the Frequency Resolution of Human Hearing

Okko Räsänen

Department of Signal Processing and Acoustics, School of Electrical Engineering, Aalto University, Espoo, Finland

okko.rasanen@aalto.fi

Abstract

The main goal of the auditory system is to detect and identify incoming sound patterns that are distributed in time and frequency. Since a priori knowledge of the spectrotemporal structure of these patterns is not available, the optimal strategy for the auditory system is to integrate incoming signals in frequency and time according to the average spectrotemporal structure of ecologically relevant stimuli. In the current work, we measure the average spectrotemporal dependencies of continuous speech and show that the dependency structure can be interpreted as an optimal filter matched to the structure of speech, and that the characteristics of the obtained filters are notably similar to the critical bands of human hearing. This result provides further evidence that speech and the auditory system are matched for optimal signaling performance and that the dependency structure is learnable with a single Hebbian-like learning mechanism.

Index Terms: speech perception, auditory perception, statistical learning, sensory plasticity

1. Introduction

Acoustic signals such as speech patterns have structure distributed in time and frequency. This means that holistic perception of these signals also requires integration of information in both of these domains across the extent of the incoming sound patterns. Signal detection theory states that the optimal signal-to-noise ratio in pattern detection under stochastic uncertainty is obtained by correlating (integrating) the incoming signal with a template of the target pattern being detected [1]. However, it is evident that the auditory system has no a priori access to the detailed characteristics of the incoming stimuli. In addition, the identification of pattern onsets from a stream of continuous and overlapping acoustic stimuli is not possible before identification of the patterns themselves, making synchronized template matching impossible.

Given these constraints, the second best strategy for a perceptual system is to optimize the signal-to-noise ratio of pattern detection across all audio stimuli that serve an ecologically relevant role for the perceiver. This means that the frequency content of the incoming audio signal is pooled across frequency channels according to the average statistical dependency between the channels. Similarly, integration in time is necessary to account for the fact that sound patterns are not localized but distributed in time, and where the average temporal structure of patterns changes with the signal frequency.

In the current work, we study the spectral and temporal dependency structure of continuous speech by using a so-called spectrotemporal dependency function (STDF), a non-logarithmic modification of the well-known mutual information function (MIF; [2,3]). Unlike the previous work [4], we extend the study of dependencies not only in time, but also across frequencies. Our statistical analysis shows that the statistical dependencies of continuous speech match well to the frequency resolution of the auditory filterbank observed in human hearing. This provides support for the idea that the human auditory system performs signal integration in a manner that optimizes the signal-to-noise ratio of pattern recognition under stochastic uncertainty.

1.1 Related work

The idea of the correspondence between the statistics of natural sounds and the properties of the human auditory system is not new. In [4] it was already shown that a single linear integrator is able to explain psychoacoustic data on different aspects of temporal auditory processing when the integrator is matched to the average temporal structure of continuous speech.

In [5], independent component analysis (ICA) was used to derive a set of auditory filters with minimal statistical dependency between the filter outputs for natural sounds. When the filter estimation was performed for 8–ms audio waveforms from both animal vocalizations and non-biological environmental sounds, it was observed that the characteristics of the learned filters have high correspondence to the physiological data observed in the human auditory system.

Recently, Ghosh et al. [6] have studied how the characteristics of an auditory filterbank are reflected in the ability to derive the underlying articulatory gestures from speech. They used mutual information (MI) to quantify the amount of dependency between articulatory gestures (from X-ray data) and the spectral representation of speech corresponding to the articulations. The spectral representation of speech was obtained from 20 brick-shaped filters that had adjustable bandwidth and center frequency and these parameters were optimized for maximal MI between the speech and articulation. What they found out was that the least uncertainty regarding the articulatory gestures was obtained for a filterbank whose center frequencies had close correspondence to the characteristics of the cochlear filterbank in the human auditory system.

As the work in [5] and [6] reveals, the auditory system seems to have adapted to the structure of natural sounds and speech. However, the existing work has not studied the dependency structure across frequencies beyond the instantaneous time scale, whereas long-term dependencies up to several hundreds of milliseconds are known to exist in natural sounds, including speech (see [7]).

In addition, the ICA-based algorithm of [5] is not biologically plausible (as already noted by the author [5]). This

leaves open the question whether the adaptation of the auditory system to the average sound structure can be based on sensory learning during early childhood (i.e., learnable by a neural substrate; cf. receptive field learning in vision [8]), or is based on natural selection. As we will see in the forthcoming sections, the current work reveals that the dependency structure can indeed be learned with a very straightforward and incremental statistical learning mechanism.

2. Methods

Unlike the ICA where the goal is to minimize the mutual information or Gaussianity of the mixing components of the original signal (e.g., [5,9]), the current approach aims to discover the characteristics of auditory filters that integrate signal in time and frequency in a manner that leads to optimal signal-to-noise ratio in the detection of patterns from continuous speech. The dependencies are measured in an abstract and non-metric state space so that the output of the analysis is scale-free with respect to the original signal (e.g., the absolute energy densities at different signal frequencies do not affect the outcome of the analysis). The outcome of the analysis is a dependency function that describes how signal content at a specific frequency is dependent on signals at different frequencies and at different temporal distances (lags), essentially being the definition of an *average pattern* in the data.

It is well known that, given a discrete sequence $X = \{a_1, a_2, \ldots, a_n\}$, $a_i \in \mathbf{A}$, the mutual information function (MIF; [2,3]) can be used to measure the mean amount of dependency (in bits) between the variables a_i and a_j separated by lag k:

$$MIF(k) = \sum_{i,j} \left(p(X(t) = a_i, X(t+k) = a_j) * \log_2 \left(\frac{p(X(t) = a_i, X(t+k) = a_j)}{p(X(t) = a_i) p(X(t+k) = a_j)} \right) \right) \quad (1)$$

If the mean temporal dependency is desired in a purely probabilistic domain, one can neglect the logarithm and compute the non-logarithmic temporal dependencies (TD) between elements:

$$TD(k) = \sum_{i,j} \frac{p_k(X(t) = a_i, X(t+k) = a_j)^2}{p(X(t) = a_i) p(X(t+k) = a_j)} - 1 \quad (2)$$

The TD essentially measures the average deviation from the statistical independence of signal events a separated by distance k, and weighted by the relative probabilities of different event pairs. The minus term is used to ensure that the TD obtains zero value in the case of purely independent variables.

However, there is no reason to limit the analysis purely to the temporal dimension, but one can also measure the dependency both across time and across multiple parallel sequences. Given that the elements of a two dimensional speech spectrogram $\mathbf{X}(t,f)$ are quantized into a finite number of partitions ($\mathbf{X}(t,f) \in [1, \ldots, N_A]$), the dependencies across the frequency channels as a function of lag can be also estimated. In this case, the spectrotemporal dependency function (STDF) obtains the form

$$STDF(f_1, f_2, k) = \sum_{i,j} \frac{p(\mathbf{X}(f_1, t) = a_i, \mathbf{X}(f_2, t+k) = a_j)^2}{p(\mathbf{X}(f_1, t) = a_i) \mathbf{X}(f_2, t+k) = a_j)} - 1 \quad (3)$$

STDF describes the mean statistical dependency between the frequencies f_1 and f_2 when f_2 is delayed by k frames with respect to f_1. If one wishes to estimate the overall dependency between two frequency channels, one can simply integrate over the lags **k** in order to obtain the spectral dependency function SDF:

$$SDF(f_1, f_2) = \sum_k STDF(f_1, f_2, k) = \sum_{i,j,k} \frac{p(\mathbf{X}(f_1, t) = a_i, \mathbf{X}(f_2, t+k) = a_j)^2}{p(\mathbf{X}(f_1, t) = a_i) p(\mathbf{X}(f_2, t+k) = a_j)} \quad (4)$$

Note that the temporal dependency structure of natural signals diminishes as k increases and therefore the integral converges to a finite value. For example, the temporal dependencies of continuous speech span to approximately 200-300 ms [7], corresponding to the time constants also observed in auditory perception [4].

In the current work, the STDF was estimated for continuous speech from two different languages. A total of one thousand randomly chosen utterances from the TIMIT (male & female train set), CAREGIVER Y2 FIN corpora [10]; 2 males, 2 females) and from an in-house phonetically balanced Finnish corpus (two males) were used in the analysis.

All signals were resampled to a sampling rate of 16 kHz before further processing. The standard magnitude spectrum of FFT was computed for all speech material using a 12–ms Hamming and a window shift of 4 ms, yielding a total of 97 frequency bins including the DC (~520 000 signal frames). Each FFT frequency bin was then treated as an individual time-series that was fed to the standard k-means algorithm in order to estimate $N_A = 8$ quantization levels for the given frequency. Finally, all data was quantized using the obtained codebooks and the STDFs and SDFs were computed from the data according to Eqs. (3-4) for all 97 frequency bins and lags **k** = {1, 2, …, 25} corresponding to the temporal delay range of 4 – 100 ms.

3. Results

Figure 1 shows examples of the estimated spectrotemporal dependency functions for point frequencies (f_1 in Eq. (3)) of 200, 1000, 4000 and 6000 Hz. As can be observed, the dependency functions resemble typical tuning curves observed in auditory nerve firing data. The average dependency across frequency increases with increasing center frequency, whereas dependency in time decreases relatively quickly from the maximum level but still has a long tail especially at low frequencies. In addition, the shape of the filter is not symmetric in the frequency domain but the slope of the high-frequency side is steeper, similarly to the tuning curves of auditory nerves in the human auditory system.

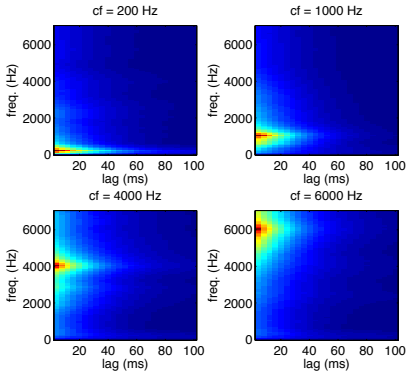

Figure 1: *STDFs for four point frequencies of 200 (top left), 1000 (top right), 4000 (bottom left) and 6000 Hz (bottom right).*

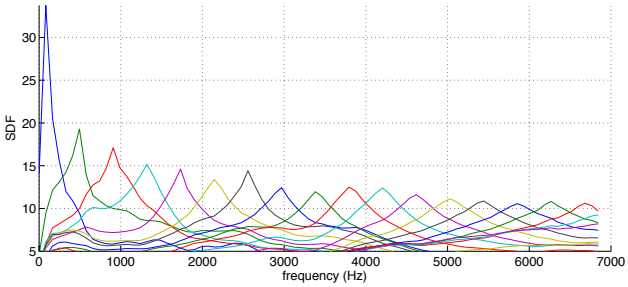

Figure 2: *SDFs obtained from continuous speech. Only the SDFs of every fifth frequency bin are shown for visual clarity.*

Figure 2 shows the SDFs (Eq. 4) for the frequency range of 0-7000 Hz. As can be again seen in the figure, the output of the analysis is notably similar with the typical auditory filter bank representations. However, the dependency does not drop to a zero level outside the main peak of the dependency function, but weak coupling can be observed across nearly 2000 Hz distances. This is mainly a result of the fact that speech is wide-band in nature, and factors contributing to the changes in the overall spectrum level (e.g., prosody) and wide-band spectra of, e.g., plosive bursts and fricatives are likely reflected in the dependency measure.

As an interesting detail, the frequency range of 0-300 Hz, the band corresponding to the typical fundamental frequency of speech, is characterized by very high spectrotemporal dependencies inside the band, whereas the dependencies to other frequencies are very low. This reveals the decoupling of glottal source from the articulatory control of the vocal tract (cf., [11]) at a statistical level and suggests that an optimal speech perception device should utilize similar type of activity pooling in the frequency range of F_0.

In order to compare the obtained filters to the physiological data on the human auditory system, bandwidths of the filters were measured. However, the SDF filters do not have a direct correspondence to the energy of physical signals, but simply represent the statistical dependencies between signal frequencies. Therefore, the lower and higher cutoff-frequencies for each filter centered at f_1 were defined as the points where the statistical dependency had attenuated $\delta = 1.5$ units from the maximal value at f_1. Since the frequency resolution of the original analysis was only 83.3 Hz, the attenuation was measured from SDF curves that had been upsampled by a factor of ten using low-pass interpolation [12]. This procedure prevents the bandwidth estimation for very low and very high frequencies due to the lack of data points, but provides a more systematic result for the remaining frequencies.

Figure 3 shows the obtained bandwidths. In addition, the physiologically motivated critical bandwidths of hearing are shown in terms of Bark [13] and equivalent rectangular bandwidth (ERB; [14]; computed according to [15]), scales. As can be seen, the correlation between the measured bandwidths and the ERB scale is very high ($r = 0.95$), meaning that the SDF essentially reconstructs the critical bands of hearing purely from the statistics of speech signals in an unsupervised and parameter-free manner. Although there is slight over-estimation of the bandwidth at low frequencies (< 2000 Hz), the filter bandwidth increases approximately linearly with the center frequency.

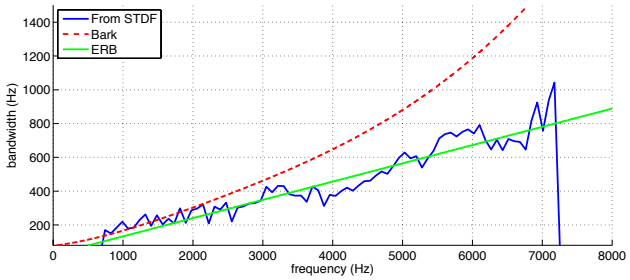

Figure 3: *Bandwidth estimates from the SDF shown in Figure 2. Equivalent rectangular bandwidths (ERB) and critical band bandwidths from the Bark scale are shown as a reference.*

As an example of spectrotemporal integration using the dependency structure of estimated from speech, Fig. 4 shows a clean speech spectrum and Fig. 5 shows a noisy speech spectrum (factory noise at SNR = 7.5 dB from NoiseX database) smoothed with the STDF filterbank (N = 97 filters). In the images, each time-frequency bin was computed as

$$\Phi(f_1,t) = \sum_{f_2}\sum_{k=0}^{K} Abs(\Im(f_2, t-k))STDF_{norm}(f_1,f_2,k)^{\alpha} \qquad (5)$$

where $\Im$ denotes Fourier-transform of the original speech signal (pre-emphasized with a standard FIR filter of form $H(z) = 1-0.95z^{-1}$). $STDF_{norm}$ is the spectrotemporal dependency function as in Eq. (3), but linearly scaled to have a minimum value of zero and a sum of one across all lags k and frequencies f. The non-linear factor α is used to control the sharpness of the filters since the STDFs are unitless and have notable proportion of the dependency mass outside the main peak region of the function. As can be observed from the figures, the STDF filtering provides subtle smoothing in time and frequency, effectively enhancing the signal-to-noise ratio for signal components that conform to the average spectrotemporal structure of continuous speech without sacrificing on the temporal accuracy of, e.g., sound onsets. Note that only one free parameter, α, is needed in the entire process of speech optimized spectrogram smoothing. Naturally, the number of filterbank outputs can be decreased to a much smaller number if compact parametrization of the data is required.

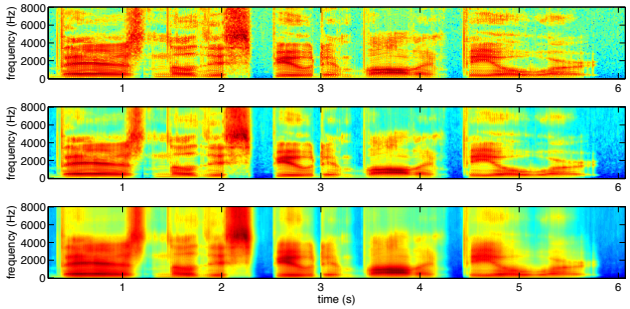

Figure 4: *Application of STDF filterbank to clean speech. Top panel: the original speech signal. Middle panel: STDF output with $\alpha = 8$. Bottom panel: STDF output with $\alpha = 4$.*

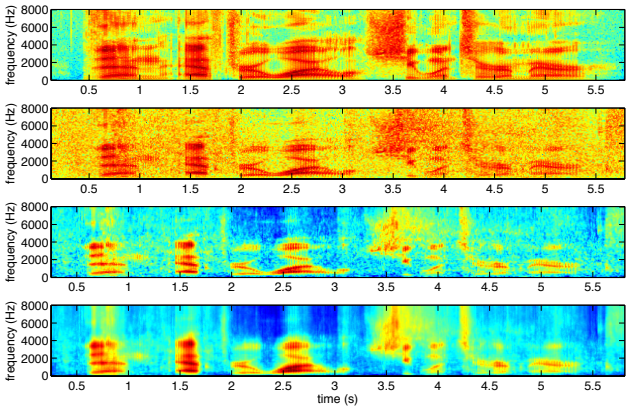

Figure 5: *Application of STDF filterbank to noised speech. Top panel: the original speech signal. Second panel: the same signal with additive factory noise at SNR = 7.5 dB. Third panel: STDF output with $\alpha = 8$. Bottom panel: STDF output with $\alpha = 4$.*

4. Discussion and conclusions

The current results show that the average statistical dependencies between signal frequencies of continuous speech match with the characteristics of the auditory filters observed in the human auditory system. Since the SDF-based filters can be interpreted as matched filters for speech, the obtained representation guarantees, on average, optimal signal-to-noise ratio in pattern detection from speech (cf. [1]). This provides further evidence that the speech and the auditory system have co-adapted for optimal signaling performance (see [5,6]).

Importantly, the current results show that the matched filters can be learned incrementally from the input data by using a simple Hebbian-like learning rule. This puts forward the question whether the critical bands of hearing are actually based on neural plasticity and auditory experience during early development, or whether they are the result of natural selection. Assuming that the physiological constraints in the inner ear and auditory nerve do not impose the given frequency resolution and integration at the middle and higher frequencies, ecologically the most efficient approach for all mammalian species would be to learn the optimal pooling of auditory information across time and frequency from every-day auditory environments. However, no definite data on early auditory capabilities of human infants are available (see [16]) since the distinction between early auditory learning and early innately specified neural development is extremely difficult.

Given the idea of a filterbank matched to the average dependencies (patterns) in signals, it is of interest whether the STDF-based filters could be used to complement the existing filterbank solutions in speech technology applications. For example, by training the STDFs directly for the patterns of interest (e.g., generic speech or specific phonetic contrasts) they can be used as theoretically optimal integrators for detection of the patterns under adverse signal conditions. Moreover, separate STDFs can be estimated for anticipated noise types, allowing the maximal contrast between the wanted signals and the noise sources. Also, the dependency measure could be used to quantify the average spectral coupling between frequency bands in telephone bandwidth extension, where earlier work has been mainly based on quantification of dependencies between Mel-scale filter outputs (e.g., [17,18]). Given the scope of the current study, these questions are left for future work.

5. Acknowledgements

This research was funded by Nokia Research Center and Nokia Foundation. The author would like to thank Unto Laine and Daniel Aalto useful comments on the current study.

6. References

[1] North, D., "Analysis of the Factors which Determine Signal/Noise Discrimination in Radar", RCA Laboratories, Princeton, N. J. Rept. PTR-6C, 1943.

[2] Li, W., "Mutual Information Functions of Natural Language Texts", Santa Fe Institute preprint SFI-89-008, 1989.

[3] Li, W., "Mutual Information Functions versus Correlation Functions", J. Statistical Physics, 60:823-837, 1990.

[4] Räsänen, O., "A unified statistical model for short-term and long-term auditory processing based on average temporal structure of signals", submitted for publication.

[5] Lewicki, M. S., "Efficient coding of natural sounds", Nature Neuroscience, 5(4):356-363, 2002.

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

[7] Räsänen, O. and Laine, U., "A method for noise-robust context-aware pattern discovery and recognition from categorical sequences", Pattern Recognition, 45:606-616, 2012.

[8] Blakemore, C. and Cooper, G., "Development of the brain depends on the visual environment", Nature, 228:477-478, 1970.

[9] Hyvärinen, A. and Oja, E., "Independent Component Analysis: Algorithms and Application", Neural Networks, 13(4-5):411-430, 2000.

[10] Altosaar, T., ten Bosch, L., Aimetti, G., Koniaris, C., Demuynck, K., & van den Heuvel, H., "A Speech Corpus for Modeling Language Acquisition: CAREGIVER", *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Malta, 1062-1068, 2010.

[11] Fant, G., "The Acoustic Theory of Speech Production", Mouton, Hague, 1960.

[12] Programs for Digital Signal Processing, IEEE Press, New York, 1979.

[13] Zwicker, E., "Subdivision of the audible frequency range into critical bands", J. Acoust. Soc. Am., 33(2):248-248, 1961.

[14] Patterson, R. D., "Auditory filter shapes derived with noise stimuli", J. Acoust. Soc. Am., 59:640-654, 1976.

[15] Glasberg, B. R. and Moore, B. C. J., "Derivation of auditory filter shapes from notched-noise data", Hearing Research, 47(1-2):103-138, 1990.

[16] Saffran, J., Werker, J. and Werner, L., "The Infant's Auditory World: Hearing, Speech and the beginnings of Language", in Handbook of Child Psychology, Vol. 2, Cognition, Perception and Language, D. Kuhn, R. Siegler, Eds., New York: Wiley, pp. 58-107, 2006.

[17] Nilsson, M., Gustafsson, H., Andersen, S. V. and Kleijn, W. B.., "Gaussian Mixture Model Based Mutual Information Estimation Between Frequency Bands in Speech", Proc. ICASSP'2002, 525-528, 2002.

[18] Pulakka, H. and Alku, P., "Bandwidth Extension of Telephone Speech Using a Neural Network and a Filter Bank Implementation for Highband Mel Spectrum", IEEE Trans. Audio, Speech, and Language Processing, 19(7):2170-2183, 2011.