SEPARATION OF SNR VIA DIMENSION EXPANSION IN A MODEL OF THE CENTRAL AUDITORY SYSTEM

Woojay Jeon and Biing-Hwang Juang

Georgia Institute of Technology
School of Electrical and Computer Engineering
Atlanta, GA 30332

ABSTRACT

In this study, we provide a theoretical approach for analyzing signal and noise separation and the noise-robustness of class-dependent activation areas in a model of the primary auditory cortex in the central auditory system. Specifically, we interpret the auditory model as a system of localized matched filters that act as a place-coding mechanism for mapping signal and noise spectra into separate regions in the three-dimensional cortical space. This framework allows us to analyze the noise robustness of signal-respondent neurons by computing their signal-to-noise ratio(SNR)'s without having to explicitly consider the complex mathematical expressions for the auditory model. The framework is also fundamentally consistent with the notion of category-dependence proposed in our previous work. Our theoretical developments of the place-coding effect and the separation of SNR will be also demonstrated experimentally.

1. INTRODUCTION

In previous studies, we experimentally established the relevance[1] of a variant of a model of the primary auditory cortex(A1) in the central auditory system[2], and were able to obtain improved recognition results under noisy conditions by introducing a phoneme category-dependent feature selection method[3] based on conjecture on the category-dependent place-coding of cognitive information. In this study, we will propose an analytical approach for studying the effects of noise on the cortical response by recognizing that the cortical transformation acts as a system of localized matched filters that map signal and noise spectra to different locations in the cortical space. The localized nature of each matched filter allows the transformation to place-code spectral components in a dimension-expanded space where they can each be isolated and accessed in a more explicit form. This is fundamentally different from the traditional cepstrum, which is a transformation that simply results in a sinusoidal decomposition of the log power spectrum.

The matched filter perspective allows us to analyze the noise robustness of cortical neurons by approximating the response areas as functions of the signal spectral envelopes, without having to directly manipulate the complex mathematical equations that model excitation and inhibition. Through this analysis, we can compute the SNR of signal-respondent cortical neurons under simplified conditions to show that the separation of spectral components allows signal-respondent cortical neurons to be robust toward noise when signal and noise are combined. These effects will also be demonstrated experimentally using samples of speech phonemes.

Furthermore, the dependence of noise robust, signal-respondent cortical locations on the structure of the frequency-domain power spectrum implies that these regions will be signal category-dependent, which is, at a conceptual level, consistent with our previous work on category-dependent feature selection[3].

2. GENERAL FRAMEWORK

2.1. Best-match response areas and place-coding

Let $p(y)$ denote the power spectrum of an uncorrupted signal defined on the frequency domain y. The cortical response $r\left(y;\lambda\right)$ represents the amount of activation of a neuron that takes on a specific neural *response area*[2]. Mathematically, the response is defined as the inner product between $p(y)$ and the response area function $w(y;\lambda)$ parameterized by λ, which consists of best frequency(BF) x, scale s, and symmetry ϕ. Since $w\left(y;\lambda\right)$ is a local response area, we assume that it is meaningful over some region $R(\lambda)$ and is zero elsewhere. We also assume that $w(y;\lambda)$ satisfies the constraint:

$$\int_{R(\lambda)} w^2\left(y;\lambda\right)dy = k \tag{1}$$

The cortical response is:

$$r\left(\lambda\right) = \int_{R(\lambda)} p\left(y\right) w\left(y;\lambda\right) dy \tag{2}$$

Assume that we are interested in the function $w(y;\lambda)$ that will provide the maximum squared (or absolute) response. By the Cauchy-Schwarz Inequality, we have:

$$r^2\left(\lambda\right) \leq k \int_{R(\lambda)} p^2\left(y\right) dy \tag{3}$$

where the maximum will occur when:

$$w\left(y;\lambda\right) = c \cdot p\left(y\right) \tag{4}$$

in $R(\lambda)$ where c is a constant designed to satisfy (1). Hence, it is generally the response area that most closely matches the shape of the spectrum (as in Fig. 3(a), (b), (d)), or its mirror (when $c < 0$ as in Fig. 3(c)) in a given local region that will result in the highest response. This was also observed in the original development of the model[2]. We can see that the cortical transformation acts like a system of localized *matched filters*[4], where each response area is designed to mimic the shape of a local spectral component.

While narrow response areas that model individual peaks give high response in harmonics of the spectrum as in Fig. 3(a), it is often response areas that match the broadband envelope of the spectrum that yield the strongest output, as in Fig. 3(b) and (c). For instance, assume that the power spectrum takes on the following form:

$$p\left(y\right) = \sum_{k\Delta \in R} \delta\left(y - k\Delta\right) v\left(y\right) \tag{5}$$

where the summation is performed over the integer k and R is the entire frequency range of interest. In a speech signal, Δ models the

pitch, while $v(y)$ is the spectral envelope that can model broadband energy distributions such as formants. We have:

$$r^2(\lambda) = \Big[\sum_{k\Delta \in R(\lambda)} v(k\Delta)\, w(k\Delta;\lambda)\Big]^2 \quad (6)$$

If Δ is small compared to $R(\lambda)$, (1) also implies:

$$\sum_{k\Delta \in R(\lambda)} w^2(k\Delta;\lambda) \approx \frac{1}{\Delta}\int_{R(\lambda)} w^2(y;\lambda)dy = \frac{K}{\Delta} \quad (7)$$

By the summation form of the Cauchy-Schwarz Inequality, the maximum response will occur when $w(k\Delta)$ is a constant multiple of $v(k\Delta)$ in $R(\lambda)$. One response area function that satisfies this is:

$$w(y;\lambda) = \begin{cases} c\cdot v(y) & y \in R(\lambda) \\ 0 & y \notin R(\lambda) \end{cases} \quad (8)$$

and we now have a response area that traces the spectral envelope.

The localized matched filtering also implies a *place-coding* mechanism in the cortical transformation. Consider the addition of some wide sense stationary noise in the time-domain that results in a corrupted spectrum written as follows:

$$p(y)' = p(y) + d(y) \quad (9)$$

Continuing our line of thought, when the input is noise alone, the cortical transformation over the region $R(\lambda)$ will be maximum for the neuron, if any, that has a response area of this form:

$$w(y;\theta) = \begin{cases} c\cdot d(y) & y \in R(\theta) = R(\lambda) \\ 0 & y \notin R(\theta) = R(\lambda) \end{cases} \quad (10)$$

Now, assume the signal power spectrum takes on the impulse train form in (5). For this signal, the best matching response area function is that given in (8). Hence, the signal and noise will each have its own distinct maximally-respondent neuron. Neurons surrounding the maximally-respondent ones will also have high responses since their response areas are similar. In summary, as long as the signal spectrum and noise spectrum are different, *the signal and noise tend to have different areas of activation in the cortical space*. For example, it can can be clearly observed in Fig. 2 $\sim$ 5 that the signal-respondent components and the noise-respondent components are mapped to distinct regions in the cortical space.

Note that the localized nature of the response area plays an important role in place-coding because it allows the response areas to replicate parts of the spectrum in a divide-and-conquer-like manner as in Fig. 3 and 5 without having to match the spectrum in its entirety. Also note that the transformation in (2) is equivalent to the Fourier transform if the response areas are sinusoids spanning R. If $p(y)$ is a log spectrum, this results in the cepstrum. However, from the matched filter perspective, the cepstrum is fundamentally different in that it is merely a sinusoidal decomposition of the power spectrum because the transformation functions are simple sinusoids spanning the entire frequency range, not localized response areas designed to collectively match the actual structure of the spectrum.

2.2. Noise-robustness

When signal and noise are combined, both the signal-respondent neuron in (8) and the noise-respondent neuron in (10) carry both signal and noise components due to the additive nature of the cortical transformation. We can show, however, that *the signal-respondent neuron is robust toward noise*. By (2) and (9), the response to the combination of signal and noise is :

$$r(\lambda)' = \int_{R(\lambda)} p(y)\, w(y;\lambda)\, dy + \int_{R(\lambda)} d(y)\, w(y;\lambda)\, dy \quad (11)$$

We define the SNR of the response as the ratio between the signal-respondent neuron's activation by the clean signal, and the distortion inflicted on the same neuron by the addition of noise. Since in the actual model the cortical response can be negative due to inhibitory regions in $w(y)$, we take the absolute value to represent response power. The motivation behind this equation is that $p(y)$ is already a measure of signal power, and viewing the cortical response as a weighted sum of the power spectrum, we want to preserve the units.

$$S_{r,\lambda} = \frac{|r(\lambda)|}{|r(\lambda)' - r(\lambda)|} = \frac{\left|\int_{R(\lambda)} p(y)\, w(y;\lambda)\, dy\right|}{\left|\int_{R(\lambda)} d(y)\, w(y;\lambda)\, dy\right|} \quad (12)$$

We can also define the SNR of the power spectrum in $R(\lambda)$ before cortical transformation.

$$S_{p,\lambda} = \frac{\int_{R(\lambda)} p(y)}{\int_{R(\lambda)} |d(y)|} \quad (13)$$

In the auditory spectrum [5] used in our physiological model, $d(y)$ can be negative, which is why we include an absolute value sign.

Now, assume that the noise is stationary white noise with variance β over R, and $p(y)$ is the Fourier power spectrum. This results in $d(y) = \beta$. Assuming the harmonic model in (5), the SNR of the noise-respondent neuron with response area defined in (10) is:

$$S_{r,\theta} = \frac{c\beta \sum_{k\Delta\in R(\lambda)} v(k\Delta)}{c\int_{R(\lambda)} \beta^2 dy} = \frac{1}{\beta V_\lambda}\sum_{k\Delta\in R(\lambda)} v(k\Delta) \quad (14)$$

where V_λ denotes the volume (length in 1-d case) of the region $R(\lambda)$. It is also easy to see that this is the SNR of the spectrum in (13):

$$S_{r,\theta} = S_{p,\lambda} \quad (15)$$

The SNR of the signal-respondent neuron with response area (8) is:

$$S_{r,\lambda} = \frac{c\sum_{k\Delta\in R(\lambda)} v^2(k\Delta)}{c\beta\int_{R(\lambda)} v(y)\, dy} \geq \frac{\frac{1}{n}\Big[\sum_{k\Delta\in R(\lambda)} v(k\Delta)\Big]^2}{\beta\int_{R(\lambda)} v(y)\, dy} \quad (16)$$

where n denotes the number of harmonic impulses in $R(\lambda)$ and we have applied the summation form of the Cauchy-Schwarz Inequality where equality holds when all $v(k\Delta)$ are equal.

If the pitch Δ is small compared to $R(\lambda)$,

$$\int_{R(\lambda)} v(y)\, dy \approx \Delta \sum_{k\Delta\in R(\lambda)} v(k\Delta) \quad (17)$$

In addition, we know that $n\Delta \approx V_\lambda$. Hence,

$$S_{r,\lambda} \geq \frac{1}{\beta V_\lambda}\sum_{k\Delta\in R(\lambda)} v(k\Delta) = S_{r,\theta} = S_{p,\lambda} \quad (18)$$

Hence, we can see that the signal-respondent neuron has an SNR that is greater than both the SNR of the noise-respondent neuron and the average SNR of the input signal in $R(\lambda)$. The same result can also be achieved if we simply assume that the peaks contain enough energy such that the envelope is a close approximation of the spectrum, i.e., $p(y) \approx v(y)$. The relation can break down if the response area does not encompass a broad range of harmonic peaks as assumed in (17), or, stated from a different perspective, if the envelope $v(y)$ is too different from the spectrum $p(y)$ in $R(\lambda)$.

To measure the collective effect of cortical transformations, we can define the overall SNR of a *set* $A = \{\lambda_i\}$ of cortical neurons, and the overall SNR of the power spectrum as:

$$S_r(A) = \frac{\sum_{\lambda_i\in A} |r(\lambda_i)|}{\sum_{\lambda_i\in A} |r(\lambda_i)' - r(\lambda_i)|},\ S_p = \frac{\int_R p(y)}{\int_R |d(y)|} \quad (19)$$

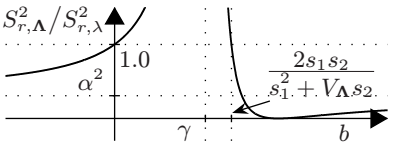


Fig. 1. Ratio of squared SNR's as a function of b

Note that S_p is simply $S_{p,\lambda}$ with $R(\lambda) = R$, and in the case of $d(y) = \beta$ denotes the overall SNR of the time-domain signal. Now, even if all neurons in A satisfied (18), this does not necessarily imply $S_r(A) \geq S_p$. However, since any lower bound on $S_{r,\lambda}$ for all $\lambda \in A$ is a lower bound for $S_r(A)$, there is a good chance of $S_r(A) \geq S_p$ as long as A is carefully selected. This turns out to be demonstrable in practice, as we will show in Section 3.

2.3. Modeling inhibition in the cortical response

The response areas in the actual cortical response are constrained to have excitatory lobes flanked by inhibitory lobes of varying scale and symmetry[1]. Therefore, (8) in the general framework can be modified to more reasonably approximate the signal-respondent response areas by adding a bias term as follows:

$$w(y;\mathbf{\Lambda}) = \begin{cases} c \cdot \{v(y) - b\} & y \in R(\mathbf{\Lambda}) \\ 0 & y \notin R(\mathbf{\Lambda}) \end{cases} \tag{20}$$

where $b > 0$ and $\mathbf{\Lambda}$ is some $\{x, s, \phi\}$. By subtracting a constant from the spectrum, we divide it into a positive region and a negative region, which are effectively matched with the excitatory and inhibitory regions of the response area. Intuitively, this makes sense because the inhibitory regions and excitatory regions tend to cancel each other, and in order to minimize this cancelation the largest spectral components should match the excitatory regions and the smallest spectral components should match the inhibitory regions as in Fig. 3(a), (b) and, (d), or vice versa as in 3(c).

To see how this affects our analysis of the SNR, we first assign the following variables for notational simplicity.

$$s_1 = \textstyle\int_{R(\mathbf{\Lambda})} v(y)\,dy, \quad s_2 = \int_{R(\mathbf{\Lambda})} v^2(y)\,dy \tag{21}$$

Again, by invoking the approximation in (17), we have:

$$S_{r,\mathbf{\Lambda}} = \frac{1}{\Delta\beta}\left|\frac{s_2 - bs_1}{s_1 - bV_{\mathbf{\Lambda}}}\right|, \quad S_{r,\lambda} = \frac{1}{\Delta\beta}\left|\frac{s_2}{s_1}\right| \tag{22}$$

We can compare the two SNR's by taking the squared ratio and writing it as a function of b as follows:

$$\frac{S^2_{r,\mathbf{\Lambda}}}{S^2_{r,\lambda}} = \left\{\frac{s_1(bs_1 - s_2)}{s_2(bV_{\mathbf{\Lambda}} - s_1)}\right\}^2 = \left\{\alpha + \frac{\rho}{b - \gamma}\right\}^2 \tag{23}$$

where

$$\alpha = \frac{s_1^2}{s_2 V_{\mathbf{\Lambda}}}, \quad \rho = \frac{s_1\left(s_1^2 - V_{\mathbf{\Lambda}} s_2\right)}{s_2 V_{\mathbf{\Lambda}}^2}, \quad \gamma = \frac{s_1}{V_{\mathbf{\Lambda}}} \tag{24}$$

By the integral form of the Cauchy-Schwarz relation, and ignoring the equality case which would require the spectral envelope to be constant, we have $s_2 > s_1^2/V_{\mathbf{\Lambda}}$. Hence, we know that $0 < \alpha < 1$ and $\rho < 0$, and also $\gamma > 0$. It is easy to visualize (23) as Fig. 1 and recognize that $S_{r,\mathbf{\Lambda}} > S_{r,\lambda}$ as long as:

$$0 < b < \frac{2s_1 s_2}{s_1^2 + V_{\mathbf{\Lambda}} s_2} \tag{25}$$

That is, the inhibitory parts can actually raise the SNR by allowing the cancelation of noise. Note that when $b = \gamma$, the noise integrated by the inhibitory part of the response area exactly cancels out the noise integrated by the excitatory part of the response area, resulting in an SNR approaching ∞ in the case of constant noise. Since γ is effectively the local average of the spectrum, it is reasonable to assume that the b for the cortical response areas, as those shown in Fig. 3, will roughly lie in the vicinity of γ due to their symmetry, particularly when $\phi = \pm\pi/2$. However, since distortion in the cortical response greatly differs from a constant, the change in SNR will not follow the illustrated curve exactly.

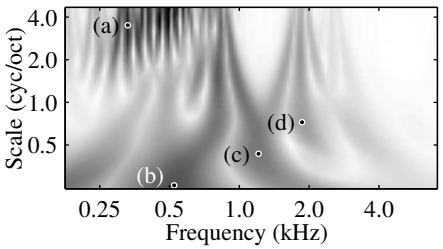


Fig. 2. $a(x, s)$ of a steady segment of an "aa" phone. Dark is high.

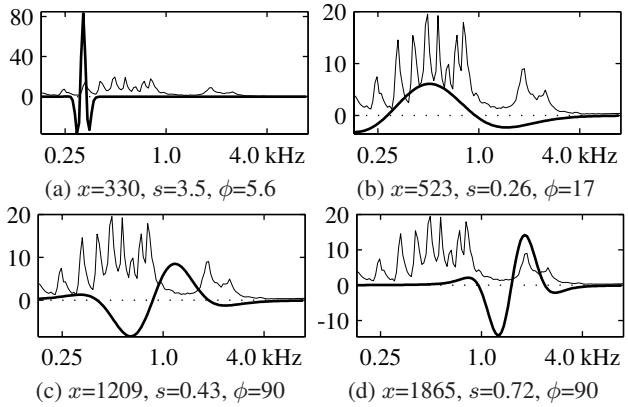


Fig. 3. The auditory spectrum of a steady "aa" phone, and response areas corresponding to components labeled in Fig. 2. Units for x, s, and ϕ are Hz, cyc/oct, and degrees, respectively. The x-axis is tonotopic frequency, and the y-axis has arbitrary units indicating the magnitude of the response areas and the auditory spectrum.

3. EXPERIMENTS

As stated in [2], it is yet unclear how the response areas in the cortical response should be normalized. If $w(y; x, s, \phi)'$ denotes the existing response areas used in the current auditory model[1], it is easy to show that (1) will be satisfied if:

$$w(y; x, s, \phi) = \frac{1}{\sqrt{\alpha^s}} w(y; x, s, \phi)' \tag{26}$$

That is, the currently-existing response areas are essentially similar to the normalized ones, only, there exists a bias that causes them to favor low-bandwidth (high s) response areas more. However, for a given scale s, which roughly translates to a fixed volume V_λ, the cortical response will behave exactly the same as when using the normalized response areas. Hence, we believe the matched filter framework essentially remains valid for the current model.

The response $a(x, s)$ is provided by the neuron with ϕ that gives the highest response for a neighborhood of x roughly defined by s:

$$a(x, s) = \max_{\phi} |r(x, s, \phi)| \tag{27}$$

Fig. 2 shows $a(x, s)$ for a steady segment of the "aa" phone. The areas with highest response constitute the signal-respondent cortical neurons. As labeled in the diagram, we can see how the harmonic (a), broadband energy (b), trough (c) (for which $c < 0$ in (20)), and formant in (d) map to separate regions. Although phase information is lost in the diagram, one can see in Fig. 3 that the signal-respondent

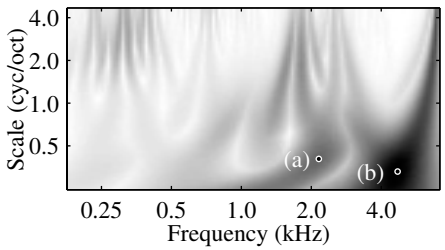
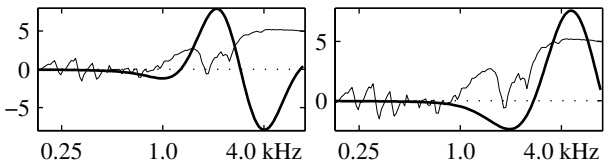


Fig. 4. $a(x, s)$ of the averaged distortion of the "aa" phone in Fig. 2 for input SNR 5 dB.



(a) x=2154, s=0.40, ϕ=-90 (b) x=4699, s=0.33, ϕ=-34

Fig. 5. Response areas of key components in Fig. 4. Units are the same as in Fig. 3. Most of the noise is mapped to cortical regions that are separate from the signal-respondent regions in Fig. 2.

clumps of neurons also have different phases, which means that they form separate clusters in the 3-d cortical space. Fig. 4 shows $a(x, s)$ of the distortion $d(y)$ for the same "aa" phone segment. In order to remove statistical variation, we computed the mean of the combined spectrum in (9) over many instances of additive white Gaussian noise in the time-domain, and then subtracted the signal spectrum $p(y)$ to obtain $d(y)$. As shown in Fig. 5, $d(y)$ is not constant and has some dependency on the signal spectrum due to the noise suppression action of the auditory spectrum[6]. We can clearly see how the noise components map to areas different from the signal-respondent areas. In particular, 4(a) does not overlap with 2(c) and 2(d) much because they have different values of ϕ. Hence, the signal-respondent areas are able to stay intact when signal and noise are combined. The SNR of signal-respondent neurons is also demonstrated for the same signal. We applied a threshold on $r(\lambda)$ to obtain a set A of signal-respondent neurons that include the major activation areas in Fig. 2. In Fig. 6, we have indicated the overall SNR's $S_r(A)$ and S_p defined in (19). To provide some sense of how the noise robustness of signal-respondent areas changes for varying best frequencies, we also plotted a localized SNR $S_r(A(x))$ where $A\,(x)$ is the set of signal-respondent neurons in A with BF x.

Finally, in Fig. 7 we computed SNR's for 44 phoneme classes in the TIMIT database for various input SNR under stationary Gaussian white noise, using all samples from training data excluding "sa" sentences. For each phoneme class, A in (19) was constructed by finding the neurons with the top 4% absolute response, averaged over all samples of the given class. U is the *entire* set of cortical neurons. The mean of the ratios $S_r(A)/S_p$ and $S_r(A)/S_r(U)$, taken over all phoneme segments, are plotted and compared to 1 to illustrate how the signal-respondent neurons generally have higher SNR.

4. CONCLUSION AND FUTURE WORK

In this study, we have analyzed the dimension expansion of the cortical transformation by approximating it as a system of localized matched filters and showed how different spectral components can match to different areas in the cortical space, allowing signal-respondent areas to be robust toward noise. We have also showed that the existence of inhibitory areas in the cortical response can sometimes

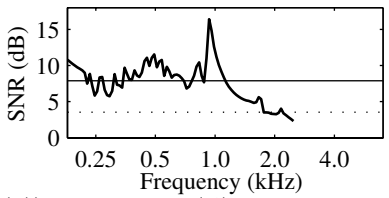


Fig. 6. $S_r(A(x))$ (thick line), $S_r(A)$ (solid horizontal line), and S_p (dotted horizontal line). $S_r(A(x))$ does not exist for $x > 3$ kHz because no signal-respondent response with best frequency in that range exists.

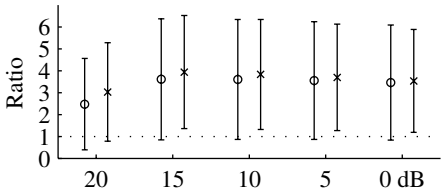


Fig. 7. Average $S_r(A)/S_p$ and $S_r(A)/S_r(U)$, marked by $\circ$ and $\times$, respectively, with error bars showing standard deviation for varying input SNR. For each input SNR, some horizontal spacing has been added between the two ratios for added visibility.

act to further boost the SNR by allowing cancelation of distortion. We demonstrated some of these effects by examples, and also verified in a preliminary experiment that the SNR of signal-respondent regions is, on average, higher than the SNR of the auditory spectrum for various samples of English phonemes.

Another important observation is that since the spectral distortion $d(y)$ in the auditory spectrum is dependent on the signal spectrum $p(y)$, and each maps to different regions in the cortical space, we can immediately conclude that noise-respondent cortical neurons, as well as signal-respondent neurons, will be phoneme class(or category)-dependent. In future work, combining category-dependent noise robustness with the category-dependent cognitive features considered in [3] could lead to better feature selection methods and improved architectures for hierarchical, category-dependent recognition and detection. We can also make better quantitative predictions on the noise separation effect in the physiological model by modeling the distortion $d(y)$ in the auditory spectrum more accurately.

5. REFERENCES

[1] W. Jeon and B.-H. Juang, "A study of auditory modeling and processing for speech signals," in *IEEE Int. Conf. Acoust., Speech. Signal Processing*, Philadelphia, PA, Mar. 2005, vol. 1, pp. 929–932.

[2] K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 5, pp. 382 – 395, Sept. 1995.

[3] W. Jeon and B.-H. Juang, "A category-dependent feature selection method for speech signals," in *INTERSPEECH-2005*, Lisbon, Portugal, Sept. 2005, pp. 365–368.

[4] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, 1993.

[5] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 824 –839, Mar. 1992.

[6] K. Wang and S. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 3, pp. 421 – 435, July 1994.