ELSEVIER

Physica D 149 (2001) 143–160

PHYSICA D

www.elsevier.com/locate/physd

A perception- and PDE-based nonlinear transformation for processing spoken words☆

Yingyong Qi[a], Jack Xin[b,*]

[a] Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ 85721, USA
[b] Department of Mathematics and TICAM, University of Texas at Austin, Austin, TX 78712, USA

Received 20 March 2000; received in revised form 31 July 2000; accepted 13 November 2000
Communicated by C.K.R.T. Jones

Abstract

Speech signals are often produced or received in the presence of noise, which is known to degrade the performance of a speech recognition system. In this paper, a perception- and PDE-based nonlinear transformation was developed to process spoken words in noisy environment. Our goal is to distinguish essential speech features and suppress noise so that the processed words are better recognized by a computer software. The nonlinear transformation was made on the spectrogram (short-term Fourier spectra) of speech signals, which reveals the signal energy distribution in time and frequency. The transformation reduces noise through time adaptation (reducing temporally slowly varying portions of spectra) and enhances spectral peaks (formants) by evolving a focusing quadratic fourth-order PDE. Short-term spectra of speech signals were initially divided into three (low, mid and high) frequency bands based on the critical bandwidth of human audition. An algorithm was developed to trace the upper and lower intensity envelopes of signal in each band. The difference between the upper and lower envelopes reflects the signal-to-noise (SNR) ratio of each band. Constant, low SNR signals in each band were adaptively decreased to reduce noise. Then evolution of the focusing PDE was used to enhance the spectral peaks, and further reduce noise interference. Numerical results on noisy spoken words indicated that the transformed spectral pattern of the spoken words was insensitive to noise for SNR ranging from 0 to 20 dB (decibel). The spectral distances between noisy words and original words decreased after the transformation. A numerical experiment was performed on 11 spoken words at SNR $= 5$ dB. A noisy word is recognized numerically by computing the closest L^2 spectral distance from the clean template. The experiment reached a recognition rate as high as 100%. Analyses on the properties of the transformation are provided. © 2001 Elsevier Science B.V. All rights reserved.

MSC: 35L60; 35R35; 41A60; 65D30

Keywords: Spoken words processing; Noise; Perception and PDE; Nonlinear transformation

1. Introduction

Speech recognition typically proceeds in two steps. First, an initial signal, an acoustic wave, is divided in time into nearly stationary overlapping frames (with a rate about one frame per 10–30 ms) and is transformed into the

☆ The work of Y. Qi and J. Xin was supported by ARO Grant DAAD-19-99-1-0248.

* Corresponding author.

E-mail addresses: yqi@u.arizona.edu (Y. Qi), jxin@math.utexas.edu (J. Xin).

PII: S0167-2789(00)00202-5

frequency space using windowed Fourier transform. The resulting spectral sequence, the spectrogram, shows the spectrum of each frame as a function of time. Hence the spectrogram is a discrete function of two variables, time and frequency. The magnitude of spectrogram represents energy (power) distribution of signals in time and frequency. In a noisy environment, the speech spectrogram contains multiple-scale information, and the essential features of speech signal (e.g. formant frequencies) are buried inside a masking noise spectrum. Thus, extracting speech spectral features and suppressing noise is one necessary step in speech recognition. Once the speech patterns are well defined, statistical tools often are used, as a second step, for pattern recognition.

Our method in this paper is focused on the first step, the signal processing step. A commonly used set of features to define the pattern of spoken words are the MFCC (Mel frequency cepstral coefficient). These coefficients arise in rational approximations in the least-square sense of the log-magnitude of the raw power spectrum with frequency scaled according to human auditory perception, the so-called Mel frequency scale (roughly logarithmic frequency scale). The scaled spectrum was decorrelated using the DCT (discrete cosine transform) to obtain the MFCC. Statistical tools, such as HMM (hidden Markov model [12]) could then be used for pattern recognition.

It is well known that noisy distortions of speech spectral data often degrade the recognition rates by creating mismatches between speech patterns of the same word. It has been well documented that preserving intrinsic speech spectral patterns and suppressing noise influence during the spectral processing stage could dramatically increase the recognition rates. In particular, incorporating time adaptation, isolating spectral peaks and enhancing peak magnitudes, help the recognition task to focus on phonetically meaningful aspects of spectrogram, and reduce the sensitivity of the pattern recognition to noise interference. Convincing experimental evidence to these effects have been demonstrated by Strope and Alwan [15], Hermansky and Morgan [7], among others.

However, the reported methods on signal processing involve many ad hoc procedures, and proceed more or less case by case. Our objective here is to introduce a systematic mathematical approach to perform adaptation and peak isolation (focusing), based on both human hearing perception and properties of a class of nonlinear PDE's (the focusing Cahn–Hillard equation [3] with quadratic nonlinearities). The end result of our treatment is a nonlinear transformation that distinguishes main speech spectral features and removes the noisy disturbances. We shall work directly on raw spectrogram and avoid making rational approximations.

Adaptation basically is to reduce any portion of a spectral curve at fixed frequency if there is not enough variation in time, a process occurring in human hearing to remove redundancy. For clean signal, the spectral curve is smooth in time, and so one can use derivative to measure this variation. Due to noise, the spectral curve becomes rough, and one must devise an alternative quantity instead. We first divide the frequencies into three (low, mid and high) bands and define an average curve for each band as a representative. Such a three band division is based on human hearing response to multi-frequency stimuli (critical bandwidth) [11]. We then construct an upper and a lower envelope for each band representative so that the difference of the two envelopes is a good measure of true signal variation in time for each band. When the difference is below a threshold (2 dB), indicating noise or redundant signal, adaptation takes place.

Peak focusing for a fixed time during voiced speech also has support from psycho-physical experiments on lateral inhibition in two simultaneous tones (see [8,9]). Auditory response to the two tones, one stronger than the other, is that the stronger tone is more pronounced and weakens the other tone. The intensity versus frequency plots of the response in [8] show the peak focusing property which is quite similar to the focusing shape in our quadratic Cahn–Hillard PDE in the sense that a peak is more focused than a nearby valley by a ratio as large as 24 dB. As pointed out in [8,9], the more pronounced peaks in auditory processing help speech discrimination and recognition both in terms of peak locations and the shapes of the peak areas.

We shall show both numerical findings and analytical properties of our transformation. Because the transformation is performed in spectral space, the adaptation, in essence, corresponds to energy dissipation, and peak focusing corresponds to scale separation in the physical space. The nonlinear processing in spectral space is different from nonlinear image processing methods on edge detection and noise removal (see [1,4,10,13] and references therein for recent literature). In image problems, nonlinear processing employs selective diffusion or smoothing and occurs directly in physical space. There are also similarities. In either application, the processing strives to enhance the desired features, such as edges in image and vision or spectral peaks in speech, and reduce noise.

The paper is organized as follows. In Section 2, we introduce our nonlinear transformation, and its numerical algorithm, and show its adaptive and focusing properties with illustrative numerics and heuristics. In Section 3, we show the original and processed spectrograms on three noisy signals with SNR equal to 5, 10, and 20 dB. The noise-free portion of the signal is the word “choice”. We also show distance plots for noisy and processed signals to the corresponding clean signals using spoken digits 1–4, and demonstrate that the processed signals are closer to the clean signals after processing than the noisy signals in the SNR range 0–20 dB. At 5 dB SNR, we show a test on 11 spoken digits with 100% rate of recognition. In Section 4, we explore the analytical aspect of peak focusing. Section 5 is the conclusion.

2. Nonlinear transformation, algorithms and properties

Let $a_{i,j} = a(t_i, x_j)$, $1 \le i \le I$, $1 \le j \le J$, be the raw speech spectral data arising from short-time window Fourier transform of a sound wave. Here I and J are positive integers, and $a_{i,j}$’s are complex numbers, however in this work, we shall only use their magnitude (intensity) $|a_{i,j}|$. The t_i’s label the nearly stationary overlapping time frames mentioned in Section 1, and x_j’s are discretizations of frequency (frequency filters) on the Bark-scale (roughly uniform in log-scale of natural frequency with unit Hz). Let the log-magnitude power spectrum be $u_{i,j} = 10(\log_{10}|a_{i,j}|^2 - \log_{10} I_0)$, where I_0 is a reference intensity. $u_{i,j}$’s are real and in unit of dB (decibel). All our computation will be conducted on u in dB unit.

For the sample of our numerical computation, $I = 43$, $J = 26$. See Fig. 1, for a greyplot of the noise-free spectrogram of the word “choice” (left frame), and the Bark-scaled log-magnitude spectrogram (right frame) that makes phonetic components clearer. In Fig. 1, the horizontal axis is time in second, and vertical axis is frequency in unit of kHz on the left frame and in unit of Bark on the right frame. We see the high frequency dark blobs during the beginning (0–0.1 s) and ending (0.4–0.55 s) periods, that correspond to “ch” and “ce”, and the two major dark blobs during the intermediate period corresponding to “oi”. Bark-scale is similar to Mel-scale, and will be used in our computation and explained more in detail in connection with critical band of human audition shortly. We shall place the $I \times J$ points uniformly on a rectangular domain $[0, T] \times [0, X]$ so that the grid size is $h \in (0, 0.09)$, $T = hI$, $X = hJ$.

When noise is present, $u_{i,j}$ receives a lot of energy towards the high frequency region, see the plot of $u_{20,j}$ for SNR $= 5$ dB (dotted line), and SNR $= 20$ dB (solid line), in Fig. 2 . Noise also makes $u_{i,j}$ rough in time.

Time and frequency directions must be treated differently, because speech signals are strongly time-dependent and the time window for processing is short. In human auditory processing, if the spectral amplitude at any fixed frequency is not varied enough over a time window, the human auditory system is going to reduce its response and ignore that segment of signal. This phenomenon is called adaptation in time and is related to the decay of an onset response of a neuron with time. The question is how we measure the variation. For fast varying rough data, it is natural to look at the distance between its upper and lower envelopes which change much less rapidly. The distance between the envelopes will encode the variation of the signal. It is unnecessary to compute envelopes for each frequency. Motivated by the critical bandwidth phenomenon in human hearing response (see [11, Chapter 4]), we divide the frequencies into three channels, low frequency channel (less than 1000 Hz), middle frequency channel (between 1000 and 1700 Hz), and high frequency channel (above 1700 Hz, often also below 5000 Hz). The low frequency channel is uniformly filtered, the filters in the middle and high frequency channels have increasing

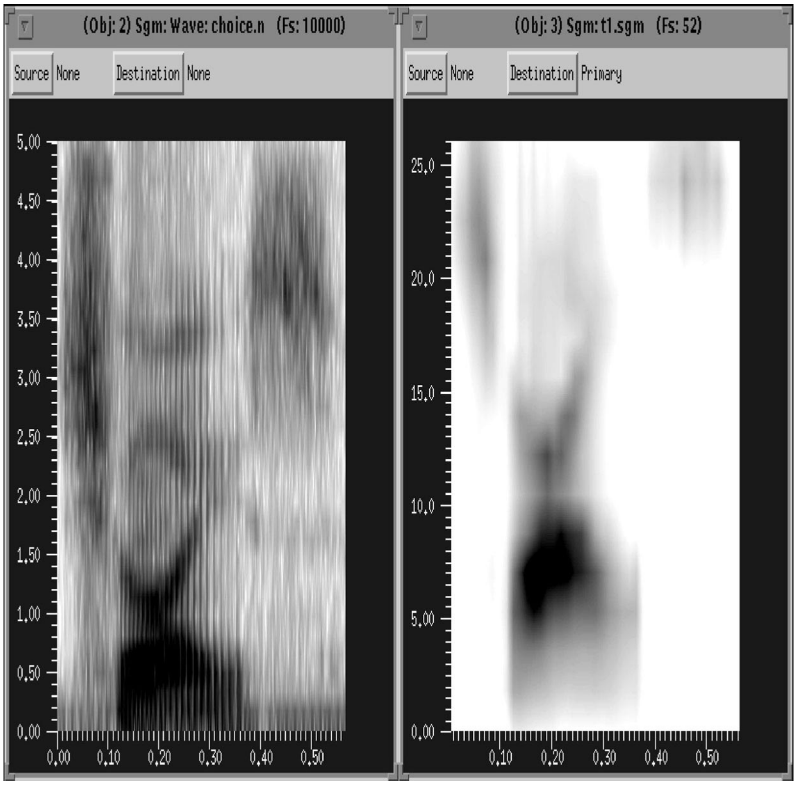


Fig. 1. Noise-free spectrogram of the word “choice”.

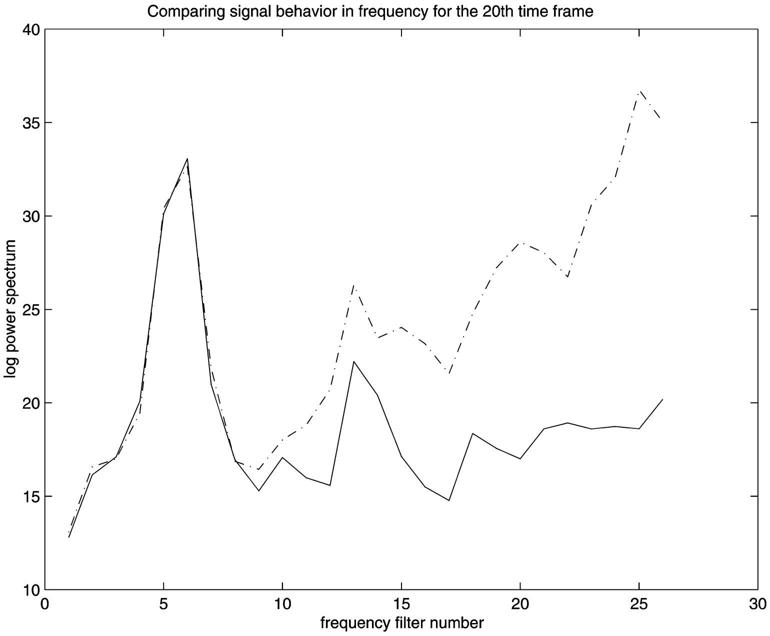


Fig. 2. Noise effect on the signal log-magnitude spectrum: energy buildup in high frequencies.

spacings in between with a growth factor of 10%. The use of critical band or the Bark-scale is similar to Mel-scale, and corresponds to constant spatial reception distances along the basilar membrane. For each channel, we compute an average curve by

$$v_{i,k} = \#\{j \in J_k\}^{-1} \sum_{j \in J_k} u_{i,j}, \tag{2.1}$$

where $k = 1, 2, 3$, and J_k is the set of filter numbers in channel k. Fig. 6 shows $v_{i,k}$ of the data signal with SNR $= 5$ dB. For all our computations in the paper, $J_1 = \{1, \ldots, 10\},\ J_2 = \{11, \ldots, 17\},\ J_3 = \{18, \ldots, 26\}$.

Next, we compute the upper and lower envelopes on $v_{k,j}$ by the following recipe. Fix any $k = 1, 2, 3$, write $v_{k,j} = v_j$ for short, and denote the upper (lower) envelopes by $\bar{e}_j\ (\underline{e}_j)$. The upper envelope is computed using local statistics, and is the sum of local mean and local deviation. The local mean is the backward five points moving average, and the local deviation is the usual empirical deviation in statistics also using five points in the past. Using past information is necessary in processing live speech signals.

Algorithm on upper envelope:

```
BEGIN
for $j = 1, \ldots, 4,\ \bar{e}_j = v_j$;
for $j = 5, \ldots, J,\ \mu_j = \frac{1}{5}\sum_{k=1}^{5} v_{j-k+1}$;
for $j = 5, \ldots, J,\ s_j = \frac{1}{4}\left(\sum_{k=1}^{5}(v_{j-k+1} - \mu_j)^2\right)^{1/2}$;
for $j = 5, \ldots, J,\ \bar{e}_j = \mu_j + s_j$;
for $j = 5, \ldots, J,\ e_{*,j} = \mu_j - s_j$;
END.
```

The lower envelope is computed based on hearing perception, instead of the $e_{*,j}$, because the distance between $\bar{e}_j$ and $e_{*,j}$ is the local deviation of the noise, and is going to overlook the imbedded signal. The lower envelope will basically follow the running minimum of $e_{*,j}$, and only follows the other parts of the $e_{*,j}$'s (increasing or decreasing but not near a running minimum) with a discount factor per $\in (0, 0.2)$. Following $e_{*,j}$ turns out to be less volatile than following v_j. However, for the first four points, $\underline{e}_j$ shall follow v_j to initiate the process.

Algorithm on lower envelope:

```
BEGIN
$e_{*,1} = v_1 = \underline{e}_1$;
STEP 1: $j = 2, \ldots, 4$,
```

$$\begin{aligned} &\underline{e}_j = \underline{e}_{j-1} + \text{per} * (v_j - v_{j-1}) && \text{if } v_j > v_{j-1}; \\ &\underline{e}_j = v_j && \text{if } v_j \le v_{j-1},\ v_j \le \underline{e}_{j-1}; \\ &\underline{e}_j = \underline{e}_{j-1} + \text{per} * (v_j - v_{j-1}) && \text{if } v_j \le v_{j-1},\ v_j > \underline{e}_{j-1}. \end{aligned} \tag{2.2}$$

```
STEP 2: $j = 5, \ldots, J$,
```

$$\begin{aligned} &\underline{e}_j = \underline{e}_{j-1} + \text{per} * (e_{*,j} - e_{*,j-1}) && \text{if } e_{*,j} > e_{*,j-1}; \\ &\underline{e}_j = e_{*,j} && \text{if } e_{*,j} \le e_{*,j-1},\ e_{*,j} \le \underline{e}_{j-1}; \\ &\underline{e}_j = \underline{e}_{j-1} + \text{per} * (e_{*,j} - e_{*,j-1}) && \text{if } e_{*,j} \le e_{*,j-1},\ e_{*,j} > \underline{e}_{j-1}. \end{aligned} \tag{2.3}$$

```
END
```

In Fig. 3, we show the computed upper and lower envelopes along with $v_{1,j}$ in channel 1, for our data “choice” with SNR $= 5$ dB. Similarly, in Figs. 4 and 5, we show the resulting envelopes in channel 2 and channel 3. The

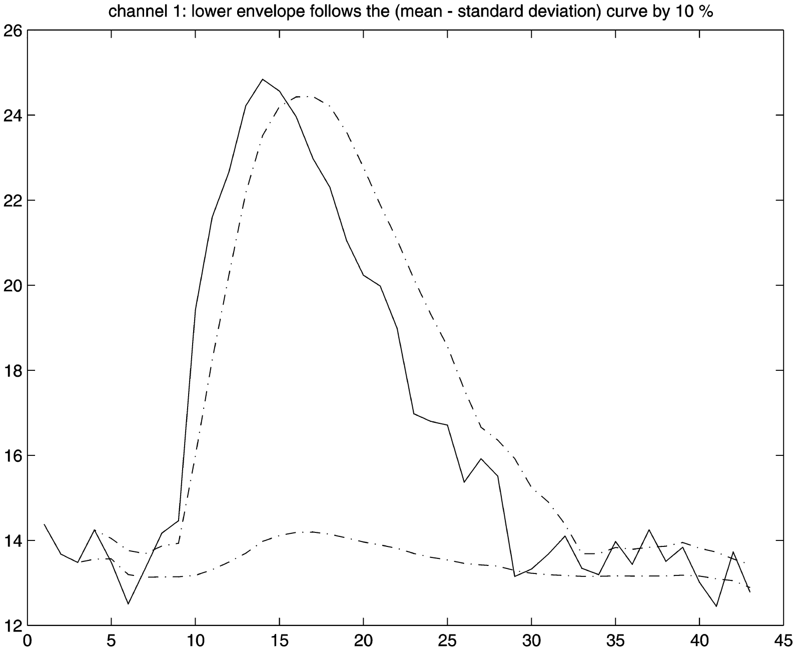


Fig. 3. Upper and lower envelopes in channel 1.

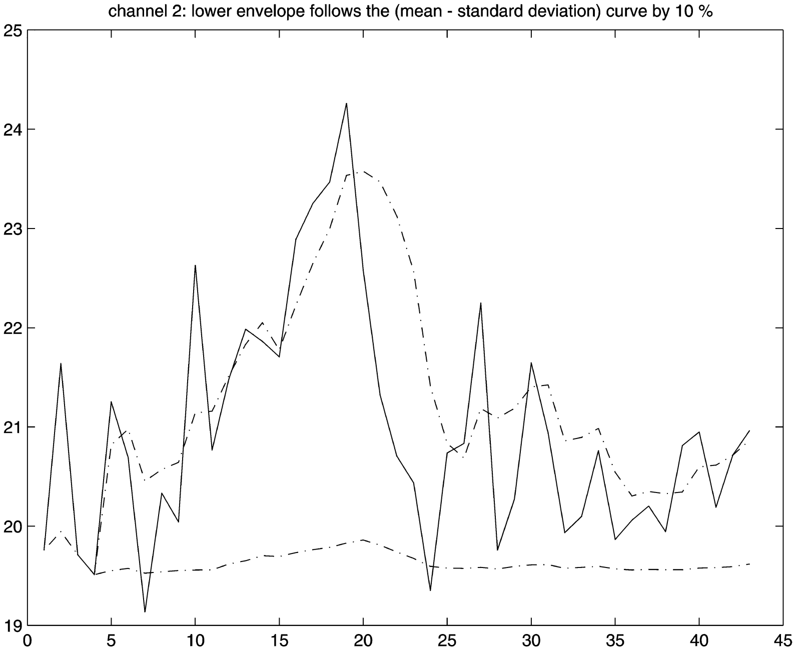


Fig. 4. Upper and lower envelopes in channel 2.

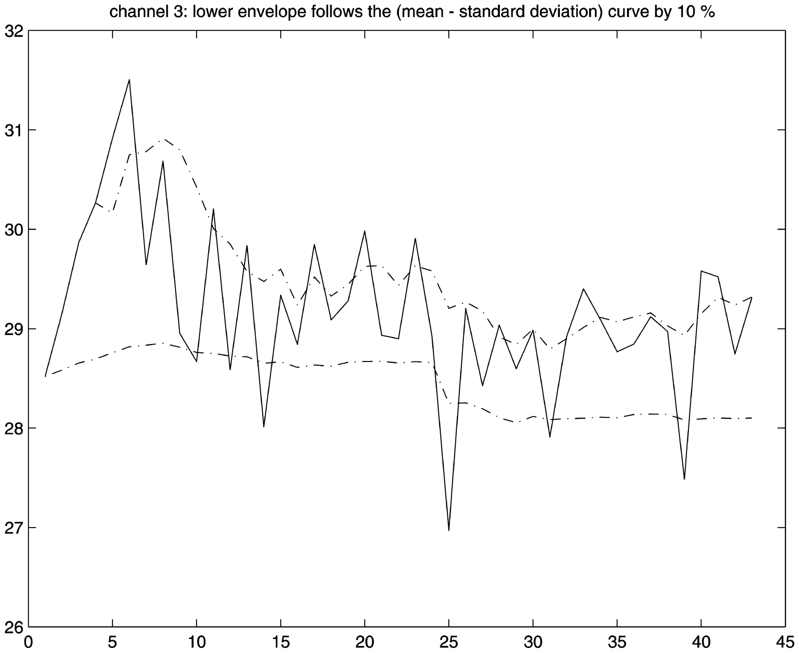


Fig. 5. Upper and lower envelopes in channel 3.

discount factor per = 10%. The upper envelopes have a slight delay response to high peaks, and lower envelopes basically follow the running minimum yet do not go as deep. The lower envelopes have much less variation than the upper envelopes, and this sets the stage for examining the difference of envelopes. In Fig. 6, we plot the three $v_{k,j}$'s, and we see that $v_{3,j} > v_{2,j}$, and $v_{2,j}$ is above $v_{1,j}$ and comparable near the peak location of $v_{1,j}$. In Fig. 7, we plot the three $\bar{e}_{k,j} - \underline{e}_{k,j} \equiv \delta_{k,j}$, and the order is reversed. The channel 1, having the first signal peak, is highest; channel 2, having the secondary signal peak, is second highest; and channel 3, with almost no major signal peaks and mostly noise, is the lowest.

We can now use $\delta_{k,j}$ as a measure of signal caused variation. If $\delta_{k,j}$ is below a threshold value in $\mathrm{dB_c}$ ($\mathrm{dB_c} =$ 2 dB in our calculation), the adaptation in time will ensue. The adaptation processing is done numerically by a discrete (e.g. finite difference) approximation of the nonlinear nonlocal PDE ($\delta(u)$ viewed as a nonlocal operator in x):

$$u_\tau + f(\delta(u) - \mathrm{dB_c}, \tau)u = 0, \tag{2.4}$$

where f is a nonnegative nonlinear function, monotone decreasing in τ, the processing time, and behaves as a transition layer (a step function) in $\delta(u) - \mathrm{dB_c}$. One simple choice of f is

$$f = \gamma\chi_{\{\delta(u)-\mathrm{dB_c}\le 0\}} \cdot \chi_{\{\tau\in[0,\tau_0]\}}, \tag{2.5}$$

where χ is the characteristic function, $\gamma > 0$, the adaptation rate, and $\tau_0 > 0$ the timescale of adaptation. Notice that if $\delta(u)$ is above the threshold $\mathrm{dB_c}$ or if $u = 0$, the solution will reach a steady state.

For small τ, solution u to (2.4) is well approximated by

$$u_0 \exp\{-\gamma\chi_{\{\delta(u_0)-\mathrm{dB_c}\le 0\}}\tau\}. \tag{2.6}$$

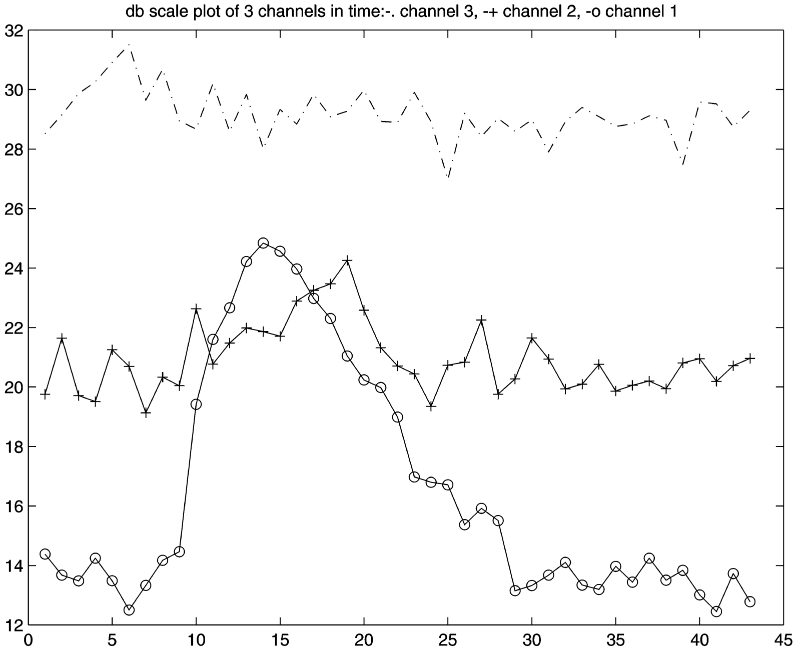


Fig. 6. The three frequency averaged curves in the three channels, respectively.

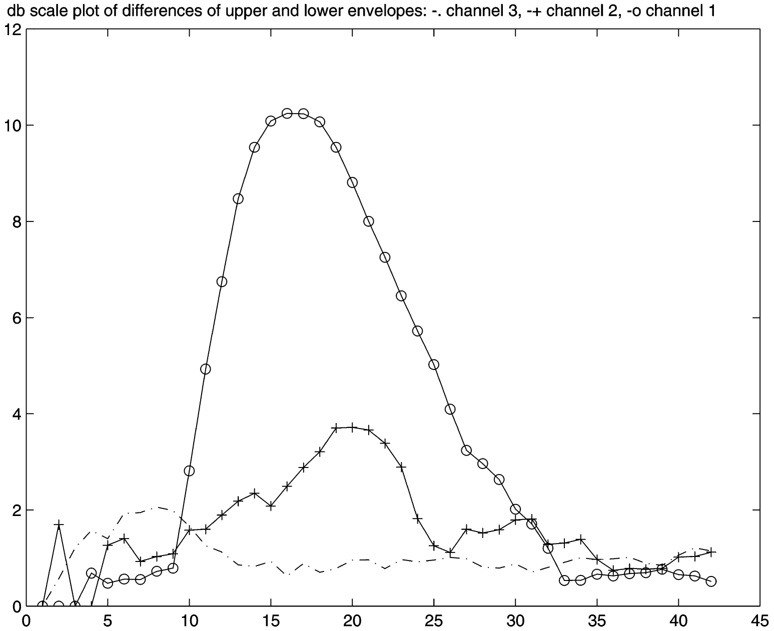


Fig. 7. Difference of upper and lower envelopes in the three channels.

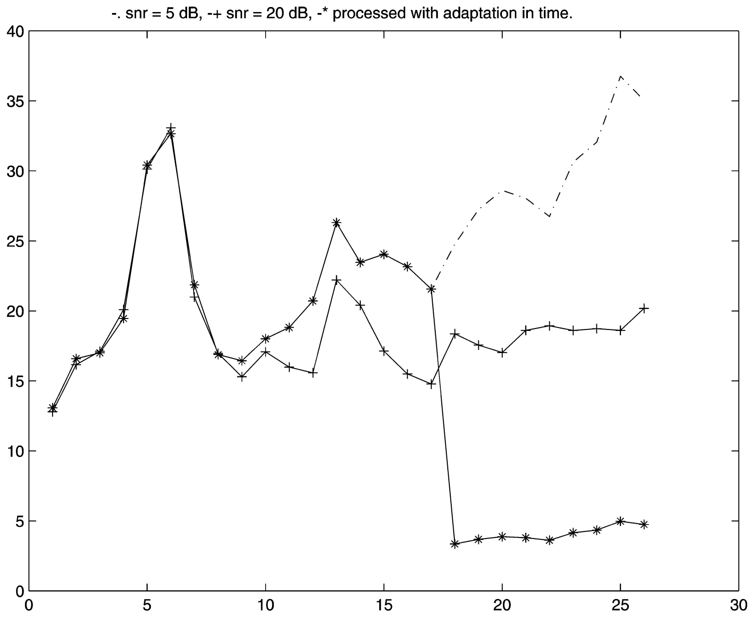


Fig. 8. Time adaptation viewed on log-amplitude/frequency plots with SNR $= 5$ and 10 dB signals.

In Fig. 8, we plot the SNR $= 5$ dB signal, its adaptation, and SNR $= 10$ dB signal against frequency filter number at the 20th time frame. The processing time is $\tau = 0.5$ in (2.6), $\mathrm{dB_c} = 2$ dB, $\gamma = 1$. We observe the noise suppression effect of adaptation. One could also take as f a smoothed version of (2.5) and the resulting adapted curve will be smoother.

Subsequent to adaptation comes the peak isolation and peak enhancement stage of processing. A PDE that is able to locally focus peaks (large curvature regions and not flat regions), and preserve L^1-norm (or energy on the log-scale magnitude), will serve the purpose. One such PDE is the following quadratic focusing Cahn–Hillard (C–H) equation:

$$u_\tau = -\alpha(u^2)_{xx} - \epsilon u_{xxxx}, \tag{2.7}$$

where $\alpha \geq 1,\ 0 < \epsilon \ll 1$. Eq. (2.7) can be rewritten as

$$u_\tau + 2\alpha u_x u_x = -2\alpha u u_{xx} - \epsilon u_{xxxx},$$

hence positive curvature regions ($u_{xx} < 0$) go up, negative curvature regions ($u_{xx} > 0$) go down. Moreover, positive slope regions advects to the right, and negative slope regions advects to the left. Combining these two effects, we have the desired local peak focusing. This motion is, however, highly unstable and requires a little bit of high wavenumber stabilization ($\epsilon > 0$) to be numerically computable, thus the C–H equation (2.7). To avoid focusing small peaks due to noise, we find it effective to first take a logarithm on initial data of u, evolve it with C–H equation, then exponentiate the result. The composition of logarithm, the C–H evolution, and exponential function on the initial data of u stabilizes the numerical solution for rough initial data, and is similar to the compression and expansion nonlinear transform on signal amplitude in the RASTA method [7].

A numerical illustration of the focusing effect is in Fig. 9, where the solid line is a smooth approximation of a fixed time voice spectral curve, and the dotted line is the evolved curve under (2.7) at time $= 0.015$, with $\alpha = 1.5,\ \epsilon = 0.09$, space step $= 0.09$, time step $= 10^{-5}$. We use second-order Crank–Nicolson implicit scheme

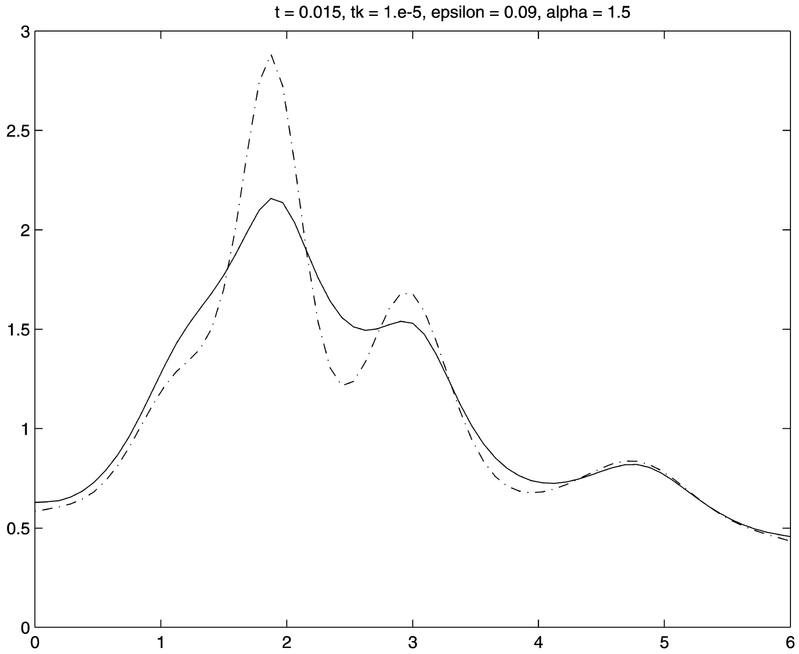


Fig. 9. Focusing effects of the quadratic C–H equation.

[14] to discretize (2.7) and impose $u_{xx} = u_{xxx} = 0$ boundary conditions at $x = 0$ and X to avoid boundary layers due to small ϵ. The computation is initiated by an explicit second-order central differencing method, and Newton iteration is carried out for each later time step.

To summarize, our nonlinear transformation consists of (1) time adaptation using perception-based three frequency channel envelopes and (2) local spectral peak focusing using the composition of logarithmic function, the C–H evolution (2.7), and exponential function. In next section, the transformation will be shown via numerical examples to yield stable results under noise perturbations, and provide useful information for recognition of noisy data.

3. Numerical results and relations to other processing methods

We shall treat the word “choice” in Fig. 1 degraded with noise of various levels, SNR = 20, 10, 5 dB. In the time adaptation stage, the threshold is 2 dB, duration is $\tau = 0.5$, $\gamma = 1$. In the peak focusing stage, spatial discretization step = 0.09, time step = 10^{-5}, and evolution takes 120 time steps (to reach a preset maximum of derivative of solution). The total time steps of evolution can be determined based on clean words as part of the training.

In Fig. 10, we show a time slice of the original signal (log-amplitude plot versus frequency) with SNR = 5 dB (solid), the adapted signal (dashed), the adapted and peak focused signal (dotted). The overall effects of processing with our nonlinear transformation can be seen on greyscale plots of original and processed signals.

In Fig. 11, the left frame is the noisy signal (SNR = 20 dB) and the right frame is the transformed signal. The horizontal axis is time, and the vertical axis is frequency. Darker colors correspond to higher energy density areas. The first two harmonics of the vowel region “oi” are enhanced and isolated. The consonant region “ch” is preserved and the energy is in the high frequency region. The “ce” consonant is downplayed a lot. Figs. 12 and 13 are obtained using the same processing parameters and stopping criterion, with noise levels (SNR = 10 dB and SNR = 5 dB).

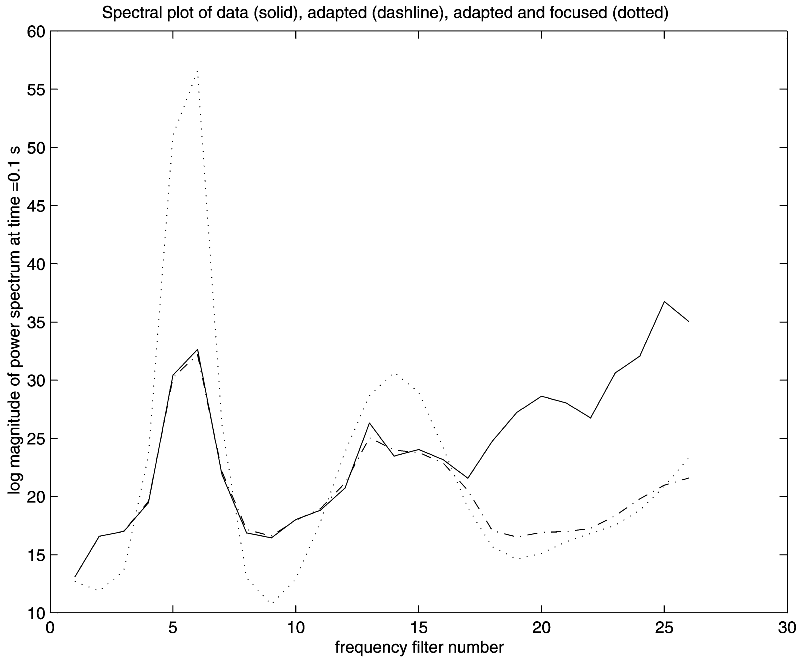


Fig. 10. Comparison of an SNR = 5 dB signal with the adapted, the adapted and peak focused signals at a fixed time 0.1 s.

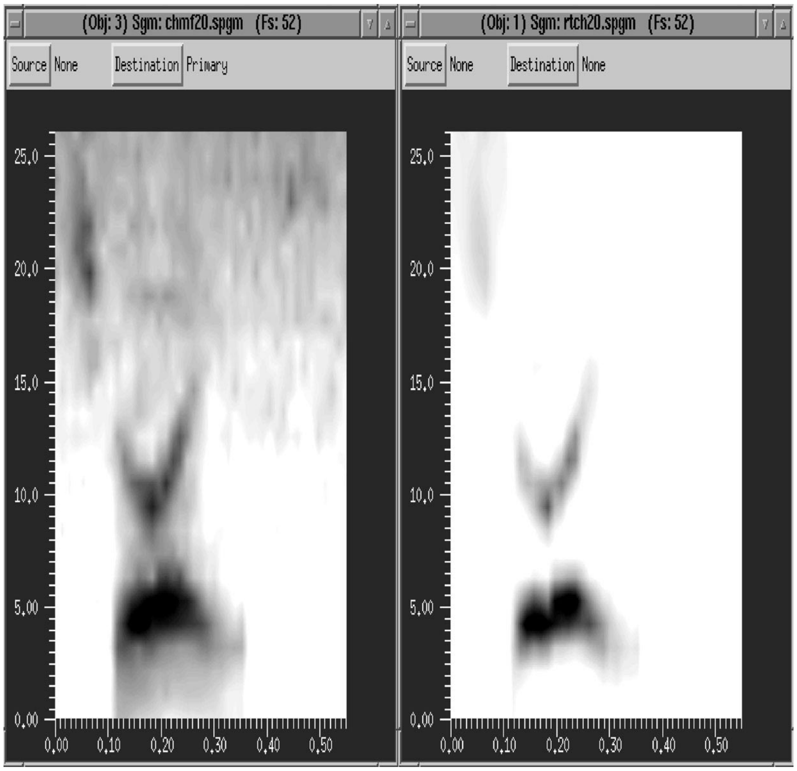


Fig. 11. Noisy (left frame) and processed (right frame) log-power spectrogram for the word “choice” with noise level SNR = 20 dB.

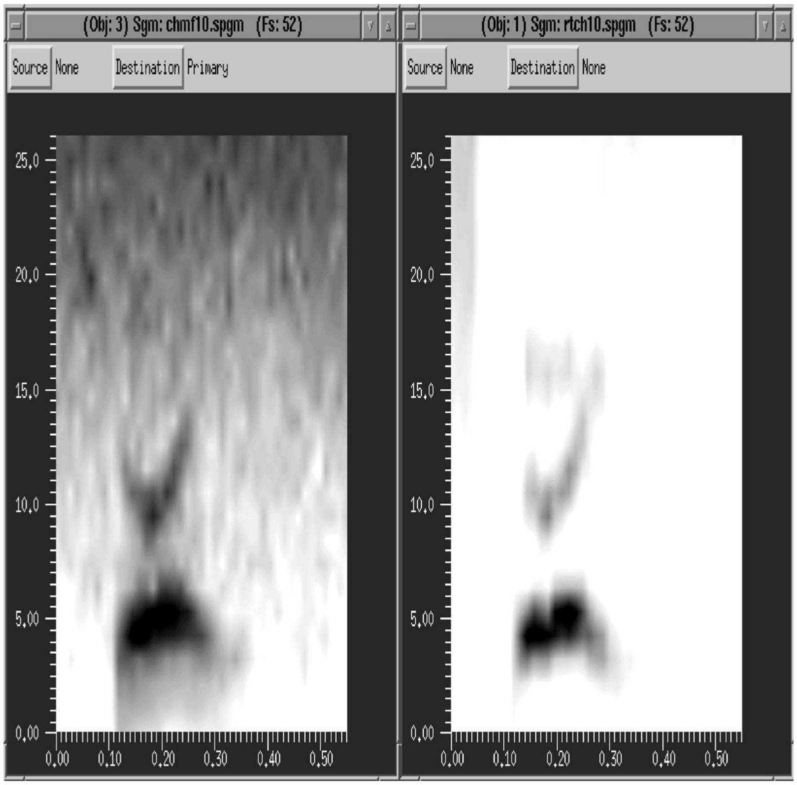


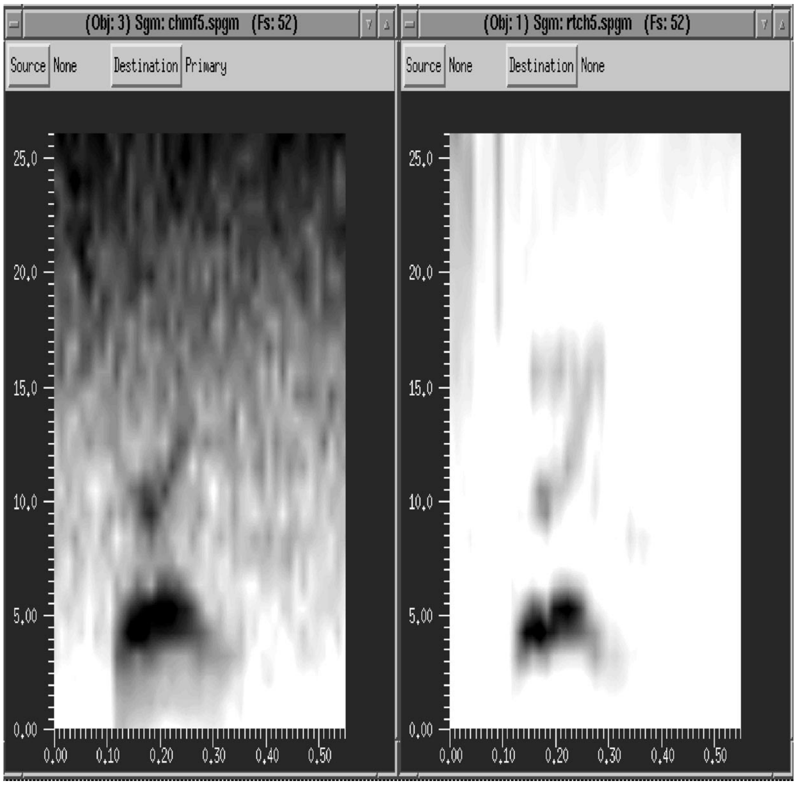
Fig. 12. Noisy (left frame) and processed (right frame) log-power spectrogram for the word “choice” with noise level SNR $= 10$ dB.

Even though the noise levels go up, the processed signals are nearly the same. This shows that the transformation is insensitive to noise perturbations, and it captures the essential features of the power spectrum.

In Fig. 14, we show the reduced L^2 spectral distances in four plots from processing noisy digits one to four with SNR range 0–20 dB, with the same parameters as before except $\alpha = 3.5$. The reduction is more pronounced for low SNR regime. Other plots (not shown) are similar for processing noisy digits five to nine, oh, and zero. The choice of other spectral L^p-norm ($p \geq 1$) [6] does not alter the results qualitatively.

In Figs. 15–17, we plot the L^2 spectral distances from noisy spoken words (one to nine, oh, zero), to the clean words at SNR $= 5$ dB. The clean words on the template have been processed with the same set of parameters as in Fig. 14. The minima of the distances select the recognized words, and we see that the processing identifies successfully all spoken words while measuring the noisy signals directly can only make out two correct words. We also notice that cross-distances tend to become smaller after processing, however, in the top left frame of Fig. 16, the distance from noisy six to clean five gets larger after processing. This is due to the focusing effect, so that the spoken word six gets penalized for missing the frequency locations of clean spoken word five. In other frames, peak focusing helps lower the distance to the correct words (say from noisy one to clean one) more than to the other words.

We also performed test for SNR $= 3$ dB, the recognition rate deteriorates to 8/11, about 73%. We carried out another test with additional noise at SNR $= 5$ dB added to the low frequencies in channel 1, the resulting rate is 10/11, about 91%. Further research is needed to improve adaptation and focusing in these settings. We also hope to exploit more of human hearing perception and find an efficient learning and training strategy.



(Obj: 3) Sgm: chmf5.spgm (Fs: 52)

Source None Destination Primary

(Obj: 1) Sgm: rtch5.spgm (Fs: 52)

Source None Destination None

Fig. 13. Noisy (left frame) and processed (right frame) log-power spectrogram for the word “choice” with noise level SNR $= 5$ dB.

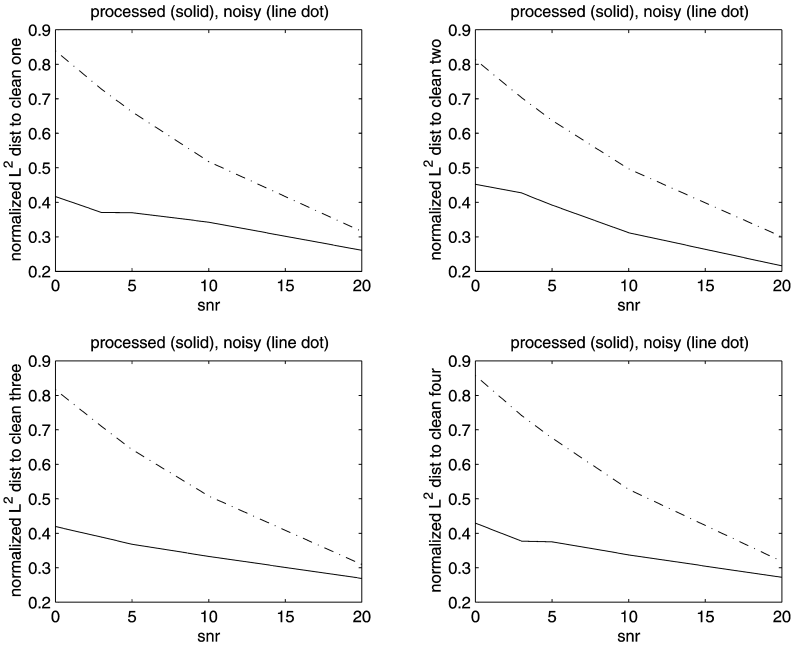


Fig. 14. L^2 distances from noisy and processed signals to clean signals (spoken digits one to four) show distance reduction after processing.

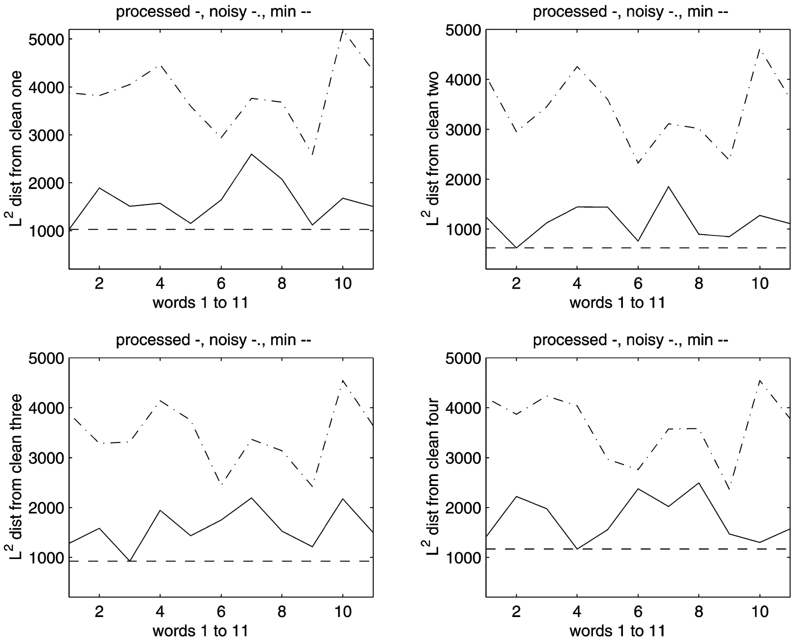


Fig. 15. L^2 distances from noisy and processed signals to clean one to four with minimum distances.

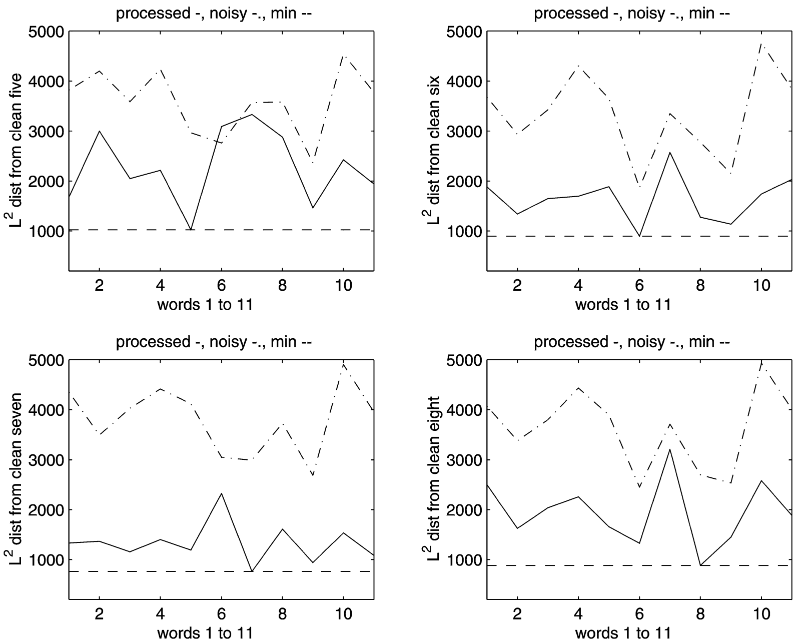


Fig. 16. L^2 distances from noisy and processed signals to clean five to eight with minimum distances.

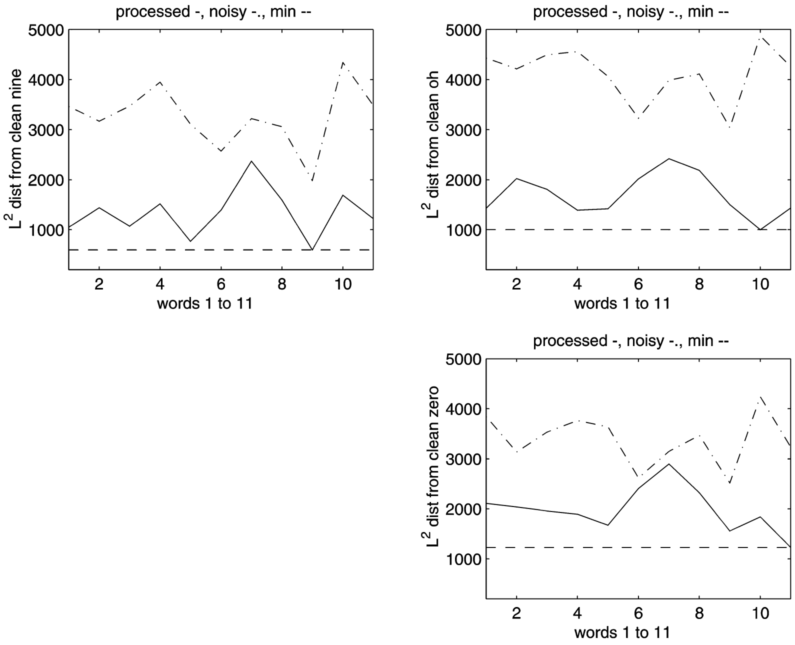


Fig. 17. L^2 distances from noisy and processed signals to clean nine, oh, zero, and minimum distances.

Let us comment on how our transformation is related to previous works. The RASTA (relative spectra) method of [7] exploits the different temporal spectral behavior of noise and signals. Time adaptation is a key component of the method in that RASTA suppressed constant or less varied portions in the time trajectory of each frequency before applying the usual rational approximation (all pole modeling). Besides this, RASTA also transforms the amplitude nonlinearly based on human hearing laws. Fig. 8 in [7] clearly shows the adaptation effects. Both time adaptation and peak focusing are utilized and advocated in [15]. It is interesting to compare our Fig. 10 with Fig. 10 in [15], and our greyplots (Figs. 11–13) with their Fig. 11, where the SNR ratios range between 5 and 40. Results of their peak isolated method (MFCCAP) resemble ours, and were found to be more robust to background noise than standard MFCC.

In the above-referenced works, the processing used estimation methods such as all pole modeling, and raised-sine cepstral liftering technique for peak isolation. The basic idea of cepstral liftering is to expand the even log-spectrum in terms of cosine basis functions, with the first few coefficients (c_0, c_1, and higher) representing log-spectrum average (c_0), log-spectrum tilt (c_1) and varying ripples (higher coefficients). Then the so-called raised-sine lifter de-emphasizes slow changes with frequency (related to overall level and vocal driving function characteristics) as well as fast changes (numerical artifacts). The processing emphasizes moderate variations with frequency, and so both spectral peaks and valleys.

Our method is based on properties of solutions to the focusing C–H equation. It treats the raw data directly without all pole approximation or DCT, and offers more flexibility and robustness in handling peak enhancement because it uses local geometrical information (curvature, slopes) without computing the expansion coefficients which also may well encode undesired features from other points (such as ripples and roughness from other frequencies away from the harmonics).

Moreover, we shall show in the next section that the C–H focusing is asymmetric in that the peak is more focused than a neighboring valley. This turns out to be consistent with the findings of Houtgast's [8] psycho-physical auditory experiment on two-tone lateral inhibition: a strong tone suppresses the nearby weak tones and there is a peak/valley height ratio from 5 to 24 dB. Houtgast attributed this phenomenon to processes within cochlear and to the nonlinear aspect of the neural coding of a stimulus sound spectrum. In a related article, Houtgast [9] pointed out that the neural projection of the actual sound spectrum near formant peaks becomes sharper, which helps vowel discrimination and recognition.

4. Analytical properties of the focusing C–H

The C–H equation (2.7) and related fourth-order parabolic equations are known to be well posed locally in time in suitable Sobolev spaces [2,5] among others. When (2.7) is considered on the interval $[0, L]$ under the zero Neumann boundary conditions: $u_x|_{x=0,L} = 0,\ u_{xxx}|_{x=0,L} = 0$, and initial condition: $u_0(x) \in H^2([0, L])$ with zero trace $u_{0,x}|_{x=0,L} = 0$, Elliot and Zheng [5] showed that if $\epsilon > L^2/\pi^2$, and u_0 is sufficiently small, then there is a global solution $u \in H^1([0,\infty); H^2([0, L]))$, so that

$$\lim_{\tau\to\infty}\left\| u - L^{-1}\int_0^L u(x)\,\mathrm{d}x \right\|_\infty = \lim_{\tau\to\infty}(\|u_x\|_\infty + \|u_{xx}\|_2) = 0.$$

This result demonstrates the smoothing effect of the fourth-order term.

If ϵ is small or initial data is not small, solutions can blow up in finite time due to the quadratic term. We are more interested in the stage prior to blow up, i.e. the focusing regime, best illustrated on the entire line by self-similar solutions, which may serve as local approximations of focusing time-dependent solutions.

With no loss of generality, set $\alpha = 1,\ \epsilon = 1$, and $u = (\tau^* - \tau)^{-1/2} f(x/(\tau^* - \tau)^{1/4})$. Then it follows from (2.7) that ($\xi = x/(\tau^* - \tau)^{1/4}$, $' = \partial/\partial\xi$):

$$\tfrac{1}{2}f + \tfrac{1}{4}\xi f' + f'''' + (f^2)'' = 0. \tag{4.1}$$

We shall only consider even solution of (4.1), and so it is enough to let $\xi \geq 0$. Eq. (4.1) has an unbounded solution $f = -\frac{1}{12}\xi^2$, on which the fourth derivative term vanishes identically. The corresponding solution $u = -\frac{1}{12}(x^2/(\tau^* - \tau))$ shows the parabolic shape of blowup profile.

Eq. (4.1) also has a first integral ([2]):

$$\tfrac{1}{4}\xi^2 f - f^2 + \xi(f^2)_\xi + \xi f_{\xi\xi\xi} - f_{\xi\xi} = c_0. \tag{4.2}$$

Integrating (4.1) over $\xi \geq 0$ indicates that any spatially decaying solution must change sign and has zero integral. Let us look for solutions to (4.1) such that $f'(0) = f'''(0) = 0,\ f(0) > 0$, and $f''(0) < 0$. Let $G = \int_0^\xi f(\xi)\,\mathrm{d}\xi$, then integrating (4.1) from zero to ξ gives

$$\tfrac{1}{4}G + \tfrac{1}{4}\xi G_\xi = -(f^2)_\xi - f_{\xi\xi\xi},$$

and using (4.2) to get

$$\tfrac{1}{4}\xi G + \tfrac{1}{4}\xi^2 G_\xi = \tfrac{1}{4}\xi^2 G_\xi - (G_\xi)^2 - G_{\xi\xi\xi} - c_0,$$

or

$$\tfrac{1}{4}\xi G + (G_\xi)^2 + G_{\xi\xi\xi} = -c_0 \equiv c_*. \tag{4.3}$$

In terms of G, any spatially decaying solution in f corresponds to a positive solution ($G(\xi) > 0$ when $\xi > 0$) of (4.3) such that $G \sim 4c_*\xi^{-1} +$ h.o.t. for large ξ.

If we regard (4.3) as an initial value problem, with initial data: $G(0) = G''(0) = 0,\ G'(0) > 0,\ G'''(0) < 0$, $c_* = (G'(0))^2 - G'''(0) > 0$, then it is not hard to show that the solutions to the initial value problem increase to a local maximum from zero then go down. How they go down to near zero depend on c_*. The solutions are oscillatory about zero if c_* is small when both $G'(0)$ and $G'''(0)$ are small in absolute value; while if c_* is large, the solutions either blow up to negative infinity in finite time or become negative with a steep negative derivative. Hence, the desired positive global solution with $G(\xi) \to 0$ can only be available for some bounded finite value of c_*. Numerical simulation of [2] showed such a solution and suggests that it exists uniquely and is dynamically attracting.

Eq. (4.3) is expedient for establishing properties of solutions. Let us first show an asymmetry property of the profile f.

Proposition 4.1. *Let* $f = f(\xi)$ *be a bounded smooth solution of* (4.1) *such that* $f(0) > 0,\ f_\xi(0) = 0,\ f_{\xi\xi}(0) \le 0,$ $f \in L^1(0,\infty),\ \int_0^\xi f(s)\,\mathrm{d}s > 0$ *for any* ξ, *then*

$$f(0) > \sup_{\xi \ge 0} - f(\xi). \tag{4.4}$$

Proof. Since f decays to zero at infinity, the supremum in (4.4) is achieved at a finite point ξ_1 where f has a negative local minimum. Eq. (4.3) reads in f:

$$f_{\xi\xi} + f^2 + \frac{\xi}{4}\int_0^\xi f(s)\,\mathrm{d}s = c_*, \tag{4.5}$$

which yields upon evaluation at both $\xi = 0$ and $\xi = \xi_1$:

$$f^2(0) \ge f_{\xi\xi}(0) + f^2(0) = f_{\xi\xi}(\xi_1) + f^2(\xi_1) + \frac{\xi_1}{4}\int_0^{\xi_1} f(s)\,\mathrm{d}s > f^2(\xi_1).$$

The proof is finished. □

Proposition 4.1 shows that the focusing tends to be more on a peak area than on a valley area. This is qualitatively similar to Fig. 9.5 of Houtgast's [8] work on lateral suppression. The assumption $\int_0^\xi f(s)\,\mathrm{d}s > 0$ is realized if f changes sign only once along $\xi > 0$, as found in its numerical computation [2].

The next property puts a constraint on $f(0)$ in terms of c^*.

Proposition 4.2. *Under the same assumptions as in Proposition* 4.1, $f^2(0) \le 6c^*$.

Proof. Multiplying (4.5) by f_ξ and integrating on $[0, \xi]$, we have

$$\frac{f_\xi^2}{2} + \frac{f^3}{3} + \frac{1}{4}\int_0^\xi \xi f_\xi \int_0^\xi f = c^*(f(\xi) - f(0)) + \frac{f^3(0)}{3}. \tag{4.6}$$

Noting that

$$\int_0^\xi \xi f_\xi \int_0^\xi f = \xi f \int_0^\xi f - \int_0^\xi f \int_0^\xi f - \int_0^\xi \xi f^2 = \xi f \int_0^\xi f - \frac{1}{2}\left(\int_0^\xi f\right)^2 - \int_0^\xi \xi f^2, \tag{4.7}$$

we have from (4.6),

$$\frac{f_\xi^2}{2} + \frac{f^3}{3} + \frac{\xi f}{4}\int_0^\xi f - \frac{1}{8}\left(\int_0^\xi f\right)^2 - \frac{1}{4}\int_0^\xi \xi f^2 = c^*(f(\xi) - f(0)) + \frac{f^3(0)}{3}. \tag{4.8}$$

Evaluating (4.8) at $\xi = \xi_1$ shows

$$\frac{f^3(\xi_1)}{3} + \frac{\xi_1 f(\xi_1)}{4}\int_0^{\xi_1} f - \frac{1}{8}\left(\int_0^{\xi_1} f\right)^2 - \frac{1}{4}\int_0^{\xi_1} \xi f^2 = c^*(f(\xi_1) - f(0)) + \frac{f^3(0)}{3},$$

and in view of Proposition 4.1,

$$c^* f(0) > \frac{f^3(\xi_1)}{3} - c^* f(\xi_1) = \frac{f^3(0)}{3} - c^* f(0) + \frac{1}{4}\int_0^{\xi_1} \xi f^2 + \frac{1}{8}\left(\int_0^{\xi_1} f\right)^2 - \frac{\xi_1 f(\xi_1)}{4}\int_0^{\xi_1} f > \frac{f^3(0)}{3} - c^* f(0). \tag{4.9}$$

It follows that $2c^* f(0) > \frac{1}{3} f^3(0)$, or $f^2(0) < 6c^*$, and the proof is complete. □

5. Conclusions

A perception- and PDE-based method is developed to process spoken words. Low, middle and high frequency channels are defined according to the critical bands of human audition. Difference of slowly varying energy envelopes is adaptively decreased in time to reduce noise. A quadratic focusing C–H equation is evolved to enhance spectral peaks corresponding to fundamental voice frequencies. The method is tested on spoken words (choice, spoken digits including zero and oh) for SNR $= 0, 3, 5, 10, 20$ dB. The L^2 spectral distances between processed signals and the corresponding noise-free signals are reduced by as much as 50% for SNR $= 0, 3, 5, 10$ dB. Numerical experiment at SNR $= 5$ dB shows 100% recognition rate on 11 spoken digits. More numerical experiments and human hearing-based modeling will be carried out for other spoken words, and at even lower SNR values.

Acknowledgements

J. Xin would like to thank Profs. L. Caffarelli, T. Hou and S. Osher for helpful discussions, and Prof. R. Diehl of Psychology Department of University of Texas at Austin for suggesting Refs. [8,9], and his encouraging remarks.

References

[1] L. Alvarez, P.-L. Lions, J.-M. Morel, Image selective smoothing and edge detection by nonlinear diffusion: II, SIAM J. Numer. Anal. 29 (3) (1992) 845–866.
[2] A. Bernoff, A. Bertozzi, Singularities in a modified Kuramoto–Sivarshinsky equation describing interface motion for phase transition, Physica D 85 (1995) 375–404.
[3] J. Cahn, J. Hillard, Free energy of a nonuniform system. I. Interfacial free energy, J. Chem. Phys. 28 (1958) 258–267.
[4] P. Blomgren, T. Chan, Color TV: total variation methods for restoration of vector-valued images, in: V. Caselle, J.-M. Morel (Eds.), PDE and Geometric Driven Diffusion in Image Processing and Analysis (special issue), IEEE Trans. Image Process. 7 (3) (1998) 304–309.
[5] C. Elliot, S. Zheng, On the Cahn–Hillard equation, Arch. Rat. Mech. Anal. 96 (1986) 339–357.
[6] R. Grey, A. Buzo, A. Gray, Y. Matsuyama, Distortion measures for speech processing, IEEE Trans. Acoust. Speech Signal Process. ASSP-28 (4) (1980) 367–376.
[7] H. Hermansky, N. Morgan, RASTA processing of speech, IEEE Trans. Speech Audio Process. 2 (4) (1994) 578–589.
[8] T. Houtgast, Lateral suppression in hearing: a psychophysical study on the ear's capability to preserve and enhance spectral contrasts, Research Monograph, Academic Press, Amsterdam, 1974.
[9] T. Houtgast, Auditory analysis of vowel-like sounds, Acustica 31 (1974) 320–324.
[10] S. Osher, L. Rudin, Feature-oriented image enhancement using shock filters, SIAM J. Numer. Anal. 27 (1990) 919–940.
[11] L. Rabiner, B.-H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993.
[12] L. Rabiner, B.-H. Juang, An Introduction to Hidden Markov Models, IEEE ASSP Magazine, 1986, pp. 4–16.
[13] L. Rudin, S. Osher, E. Fatemi, Nonlinear total variation-based noise removal algorithms, Physica D 60 (1992) 259–268.
[14] J. Strikwerda, Finite Difference Schemes and PDE, Mathematical Series, Wadsworth, Belmont, CA, 1989.
[15] B. Strope, A. Alwan, A model of dynamic auditory perception and its application to robust word recognition, IEEE Trans. Speech Audio Process. 5 (5) (1997) 451–464.