Towards Fusion of Feature Extraction and Acoustic Model Training: A Top Down Process for Robust Speech Recognition

Authors: Yu-Hsiang Bosco Chiu,
Bhiksha Raj and Richard M. Stern

Professor:陳嘉平
Reporter: 吳柏鋒

Outline

- Introduction
- Feature computation
- Learning the Non-linearity
- Estimating Sigmoidal Parameters
- Experimental Results

Introduction

- This paper presents a strategy to learn physiologically motivated components in a feature computation module and use a set of logistic functions which represent the rate-level nonlinearity

- The parameters of these rate-level functions are estimated to maximize the a posteriori probability of the correct class in the training data.

Feature computation

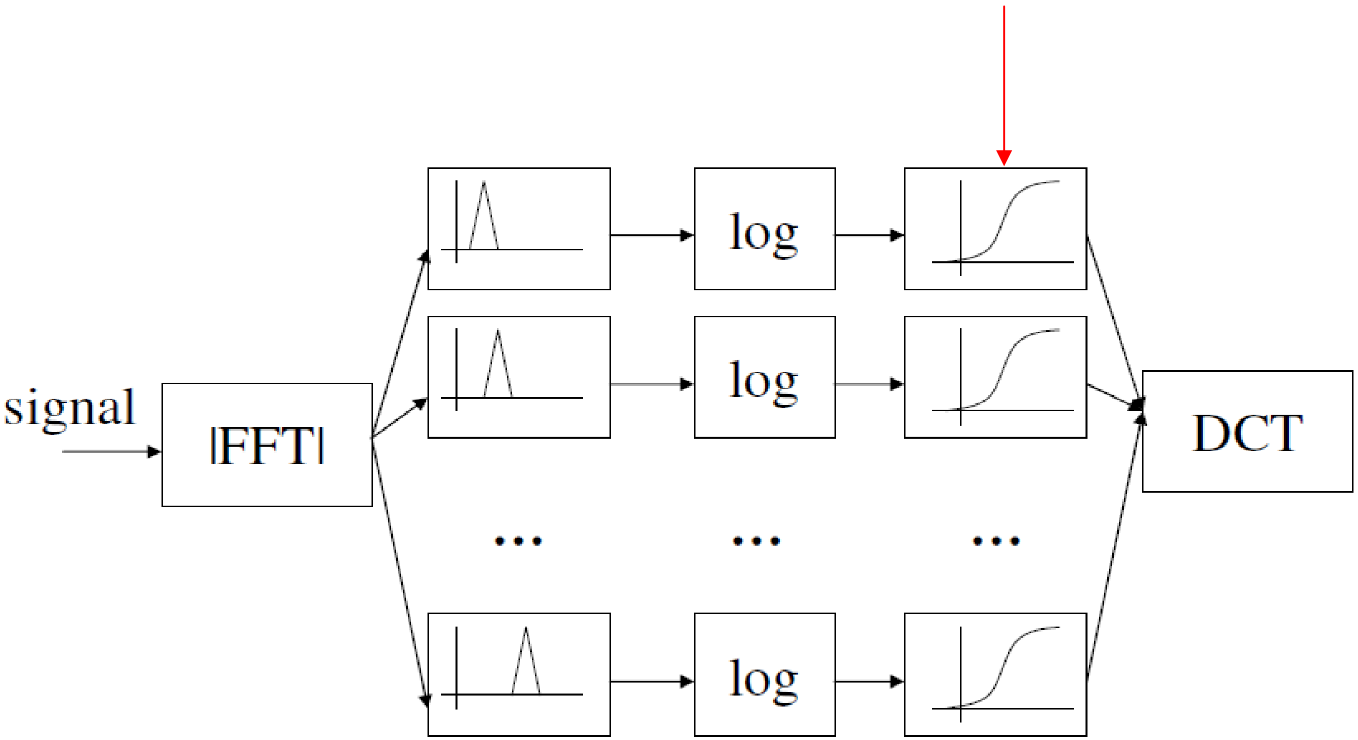
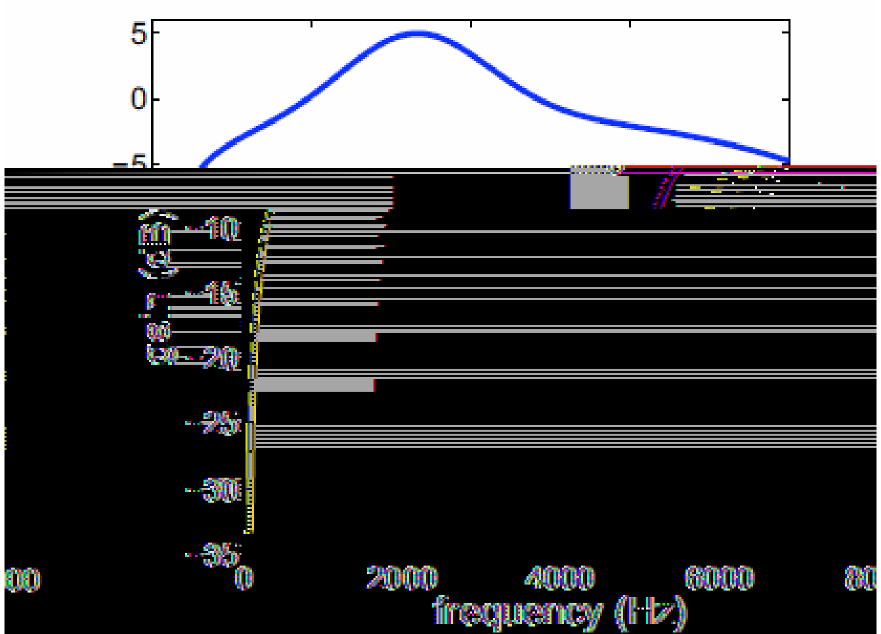


Figure 1: *Feature computation scheme*

Feature computation

- An addition aspect of the feature computation that is not illustrated in Figure 1. is an <u>equal-loudness weighting</u> that is applied to every spectral component prior to the logarithmic compression



ing. Figure 2: *Equal loudness weight*

Feature computation

- log compressed Mel-spectral values are passed through a sigmoidal nonlinearity that represents the rate-level nonlinearity

$$x_i[t] = \frac{\alpha[i]}{1 + exp(\, w_1[i] \cdot y_i[t] + w_0[i]\,)} \qquad (1)$$

where $y_i\,[t]$ is the i^{th} log Mel-spectral value, $x_i\,[t]$ is the corresponding sigmoid-compressed value of frame t. $\alpha[i] = 0.05, w_0\,[i] = 0.613, w_1\,[i] = -0.521$ $\forall i$ were obtained by fitting it to physiological measurements followed by further hand refinement

Feature computation

- In the absence of the sigmoidal nonlinearity, equal-loudness weighting emerges as an additive constant after the logarithmic compression and would get eliminated by the cepstral mean subtraction (CMS) that is routinely used in speech recognition.

- The sigmoidal non-linearity serves to combine the gain into the features in a non-linear manner such that it cannot be eliminated by CMS.

Learning the Non-linearity

- The posterior probabilities of any sound class C, Given a specific observation s is given by

$$P(C|s)=\frac{P(s|C)P(C)}{\sum_{C'} P(s|C')P(C')}=\frac{P(s|C)}{\sum_{C'} P(s|C')}$$

$$=\frac{N(s|\mu_C,\ \sigma_C)}{\sum_{C'} N(s|\mu_{C'},\ \sigma_{C'})}$$

μ_C is mean vector, σ_C is the covariance of the feature vectors for any sound class C

Learning the Non-linearity

- accumulated posterior probability of the entire training data

$$P = \prod_{u,t} \frac{N(s_{u,t} | \mu_{C_{u,t}}, \sigma_{C_{u,t}})}{\sum_{C} N(s_{u,t} | \mu_{C}, \sigma_{C})} \qquad (3)$$

$s_{u,t}$ is the feature vector obtained for the t^{th} analysis frame of the utterance u, $C_{u,t}$ is the sound class that the corresponding segment of speech

Learning the Non-linearity

- Estimating sound class distribution parameter

where $I(s \in C)$ is an indicator function that takes a value of 1 if s belongs to sound class C and 0 otherwise.

Estimating Sigmoidal Parameters

The parameters for the logistic function F={α, ω_0, ω_1} are estimated to maximize log(P) using a gradient descent approach

$$\alpha^{\text{new}} = \alpha^{\text{old}} + 0.00005 \frac{\partial \log \mathrm{P}}{\partial \alpha}$$

$$\omega_0^{\text{new}} = \omega_0^{\text{old}} + 0.05 \frac{\partial \log \mathrm{P}}{\partial \omega_0}$$

$$\omega_1^{\text{new}} = \omega_1^{\text{old}} + 0.01 \frac{\partial \log \mathrm{P}}{\partial \omega_1} \tag{5}$$

Estimating Sigmoidal Parameters

Input: F, $\{(y_{u,t}, C_{u,t}), u = 1..U, t = 1..T_U\}$

Output: F

while not converged do

1 Compute feature vector $\{s_{1,1}, \ldots, s_{U,T_U}\}$ using Eq.(1) and DCT with CMS

2 Estimate $\{\mu_{C}, \sigma_C\}$ $\forall C$ using Eq.(4) on clean training set

3 Compute log(P) using Eq.(3) on both clean and noisy training set

4 $F_{new} \leftarrow F_{old} + \frac{\partial \log P}{\partial F}$ using Eq.(5) on both clean and noisy training set

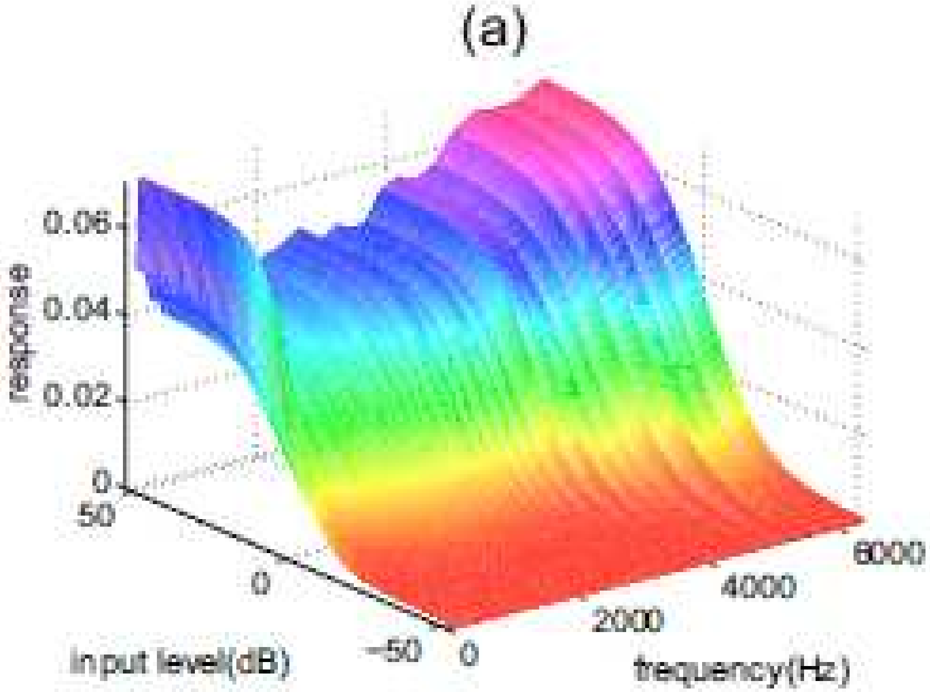
Experimental Results

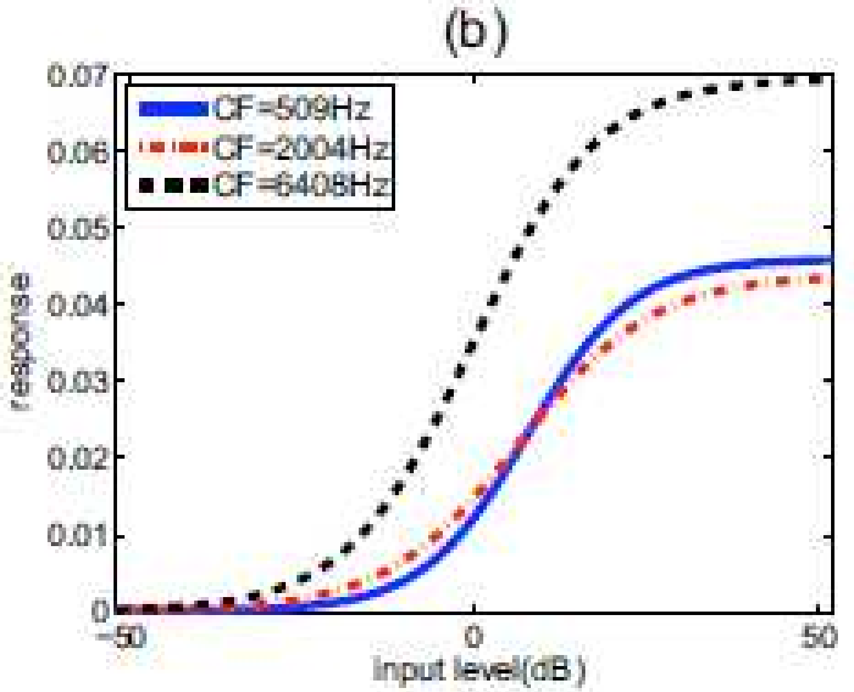
- Use the DARPA Resource Management database to evaluate the proposed method.

- In order to train the rate-level nonlinearity, the pink noise from NOISEX-92 was artificially added in to the original clean training set at 10dB SNR to create the noisy training set.

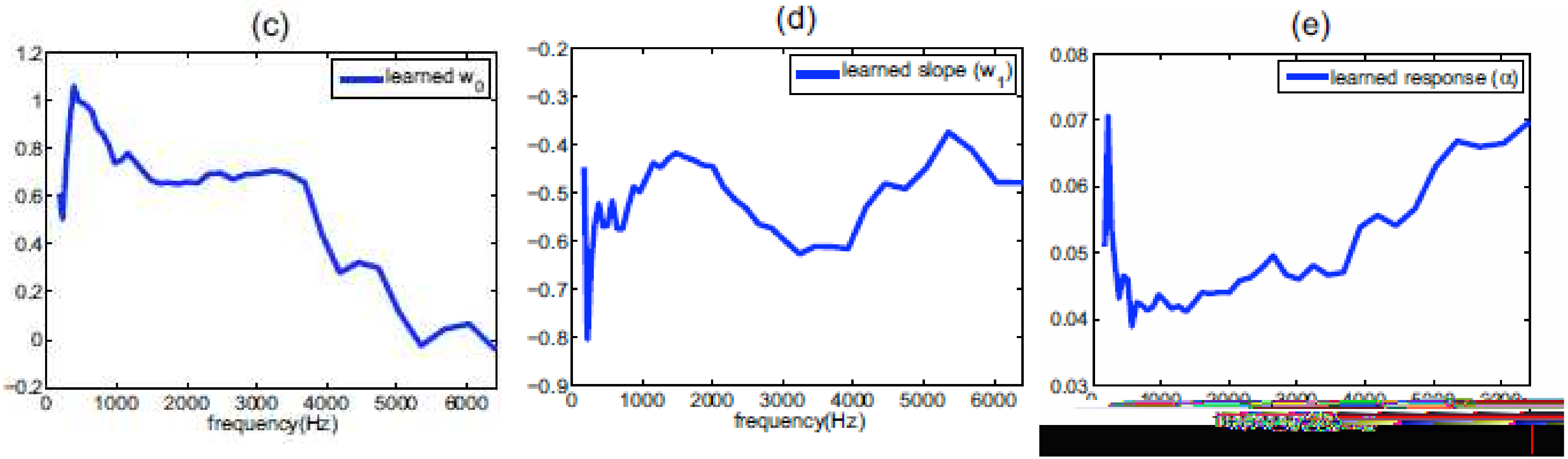
Experimental Results

- shows the rate-level nonlinearities learned





Experimental Results



Experimental Results