



Robust Speaker Recognition using Prosodic Information

Present by : 陳子和

Advisor : 廖元甫 教授



Content:

- **Introduction:**
- **Latent-Semantic Analysis-based speaker recognition:**
- **Quantizing the prosodic feature:**
- **Prosodic pattern extraction:**
- **Calculating the occurrences statistic:**
- **Applying the SVD:**
- **Measuring the score of the EPA and the system fusion:**
- **Experiment Result:**
- **Conclusions:**



Introduction:

Most speaker recognition systems utilize speaker features by looking at short-term spectral information and ignore long-term information, such as prosody and speaking style. We presents a method called eigen-prosody analysis that uses the prosodic information to capture long-term information for speaker recognition task. In experiments, even in very few training data and mismatch channel environment, a remarkable recognition rate was obtained.



Prosodic information has been applied in three main ways

- global statistics of some prosodic-based feature are compared between two utterances
- comparing the temporal trajectory of the prosodic contours.
- using n-gram language model to model the prosodic information text-independent speaker verification tasks

we present an Eigen-Prosodic Analysis approach (called EPA), which partially addresses these two questions that demonstrate effective ways to model and apply conceptual dynamic prosodic information for text-independent speaker recognition tasks.

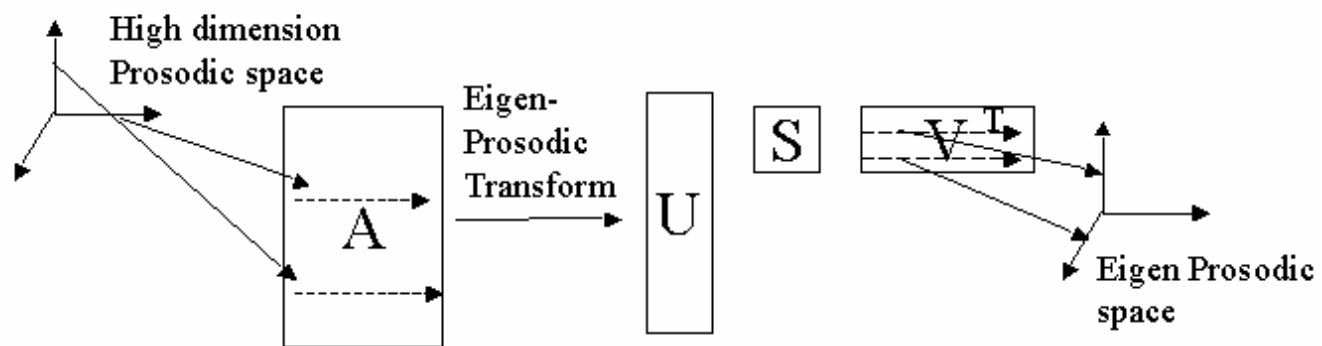
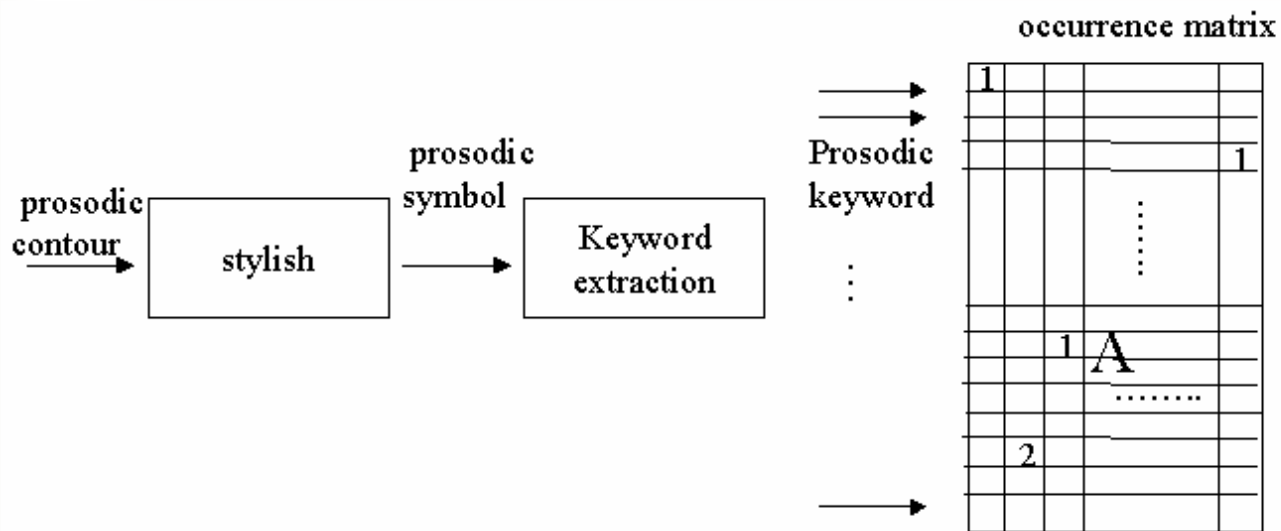


the system fusion

$$L(\theta) = \frac{1}{N_u} \sum_{u=1}^{N_u} L(\theta) = \frac{1}{N_u} \sum_{u=1}^{N_u} l(d(X_u))$$

X_u is vector of the parameter of testing score such as GMM and the EPA score

$d(\cdot)$ is distance measure for misclassifications; and $l(\cdot)$ is a soft error-counting function



Quantizing the prosodic feature:

1. pitch slop is stylized into 3 levels:
Rising (‘/’), **Falling** (‘\’), and **Flatting** (‘-‘)
2. the pause duration is quantized into 3 levels:
Short (S), **Medium (M)**, and **Long (L)**.

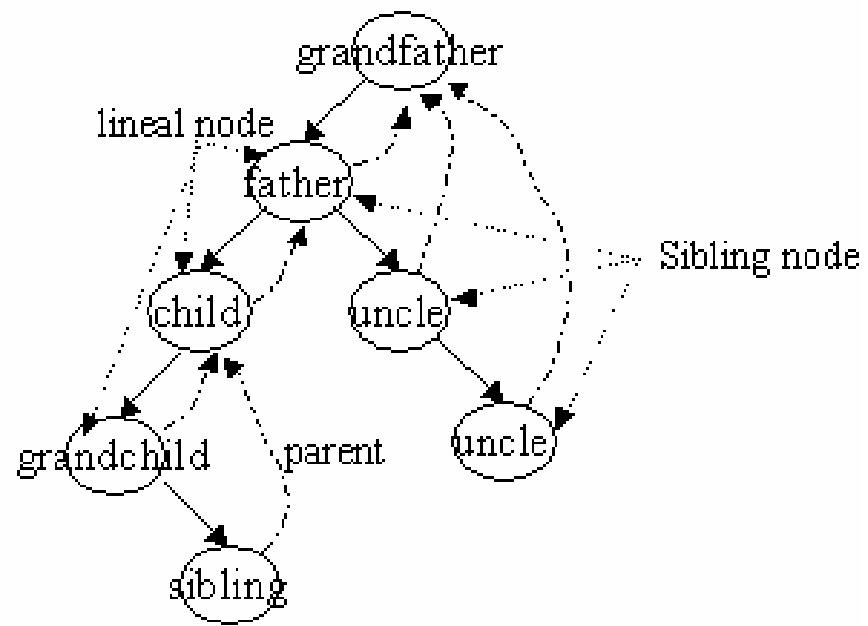


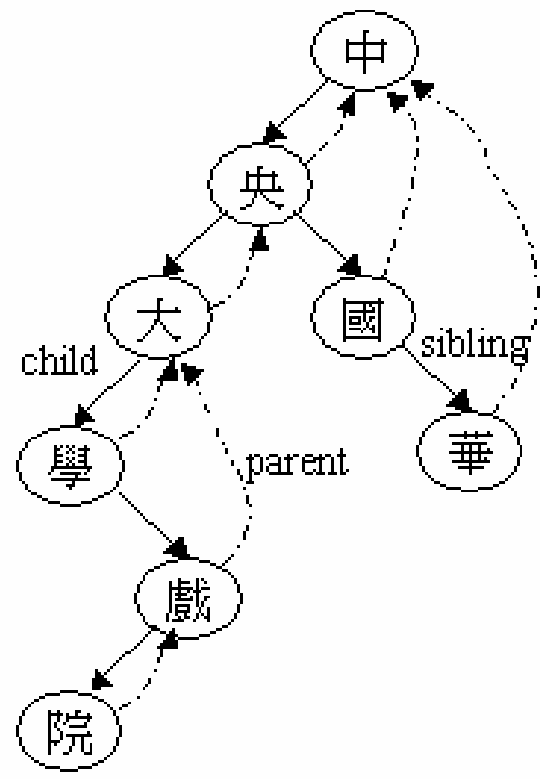


Prosodic pattern extraction:

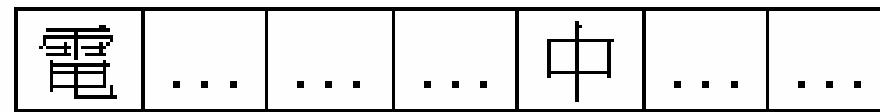
- statistical methods
- rule-base methods

We try to use entropy extraction methods as the criterion to extract the prosodic pattern.

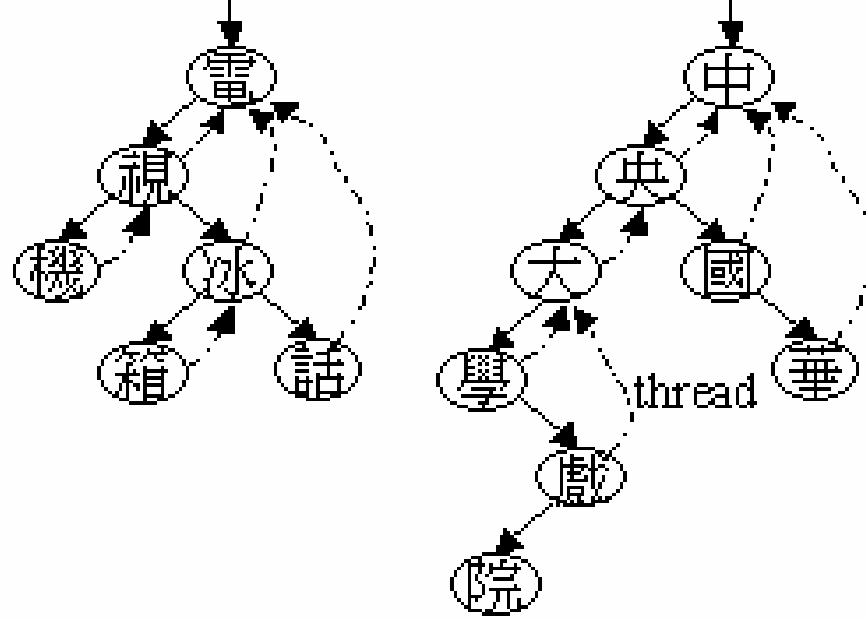


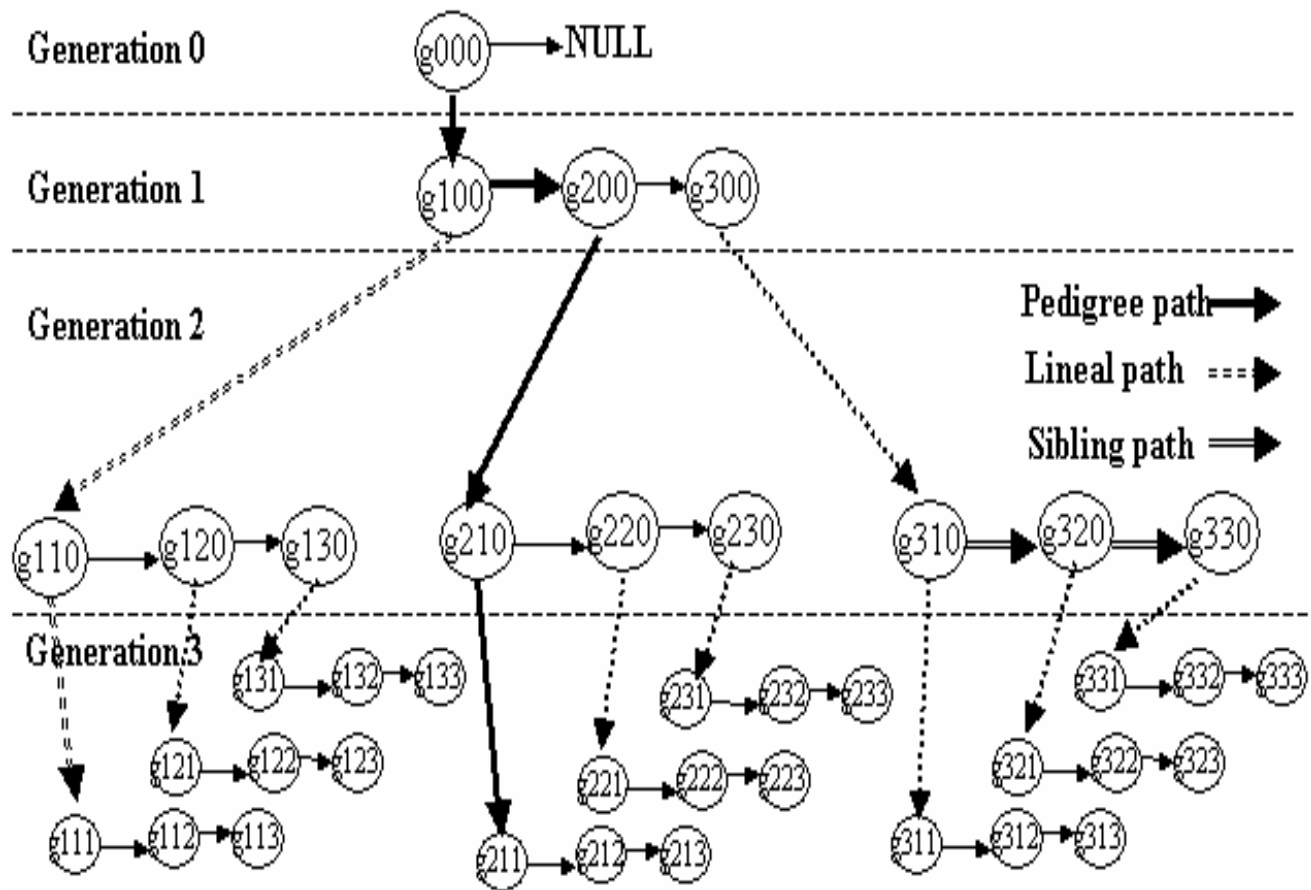


5401

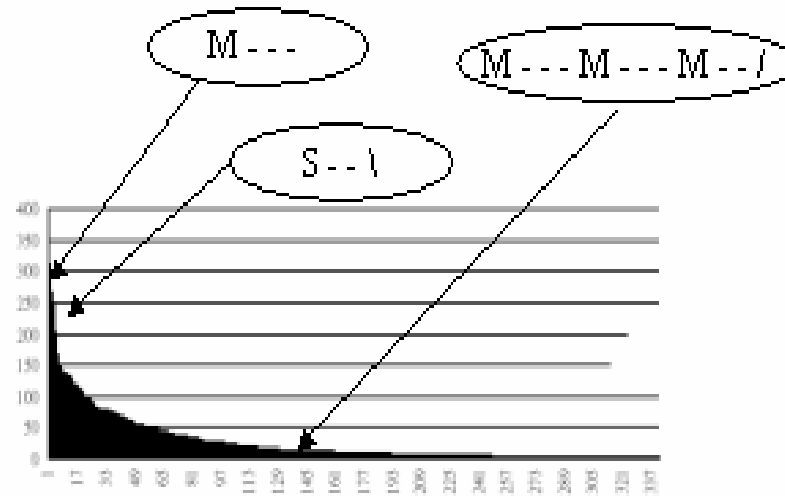


W-array






the histogram of the prosodic pattern.



The high frequency prosodic patterns such as “M---”, “S--\|” are most popular of the speakers saying and the low frequency prosodic patterns just only few person has say this pattern.



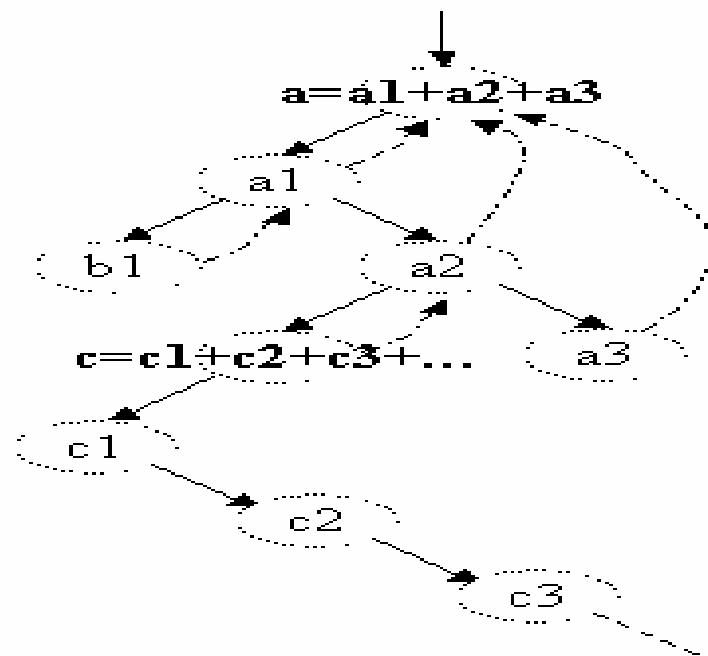
The entropy information in generation-forest is calculated by the counters of the sibling path and the results are recorded on the parent node of the sibling nodes in the sibling path. According to the entropy information, it offers a good decision criterion to estimate the keyword termination. The entropy function is defined as

$$H (X) = E \{ I (x_j) \} = - \sum_j^n p (x_j) \log_a p (x_j)$$

$$I (x_j) = \log_a \frac{1}{p (x_j)} = - \log_a p (x_j)$$

x_j be an event that occurs with probability $p(x_j)$

$p(x_j)$ denotes the counts occurring in sibling node divided by the counts occurring in its parent node





Calculating the occurrences statistic

- BIN
- IDF
- TF



Applying the SVD:

$$A = U \Sigma V^T$$

$$U = (u_1, \dots, u_m)$$

$$V = (v_1, \dots, v_n)$$

$$\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$$

Reduce Dimension

$$A = U \Sigma V^T \approx \hat{A} = \hat{U} \hat{\Sigma} \hat{V}^T$$

EPA models the long-term spectral feature and robust against the perturbations, which is resulted from prosodic pattern quantization error, by using rank reduced SVD model



Measuring the score of the EPA:

$$v_h = y_h^T \hat{U} \hat{\Sigma}^{-1}$$

y_h^T testing prosodic pattern mapping to the prosodic keywords


The score is measured by the cosine of the angle between v_h and \hat{v}

$$d(\hat{v}, h_t) = \cos(\hat{v}, v_h)$$

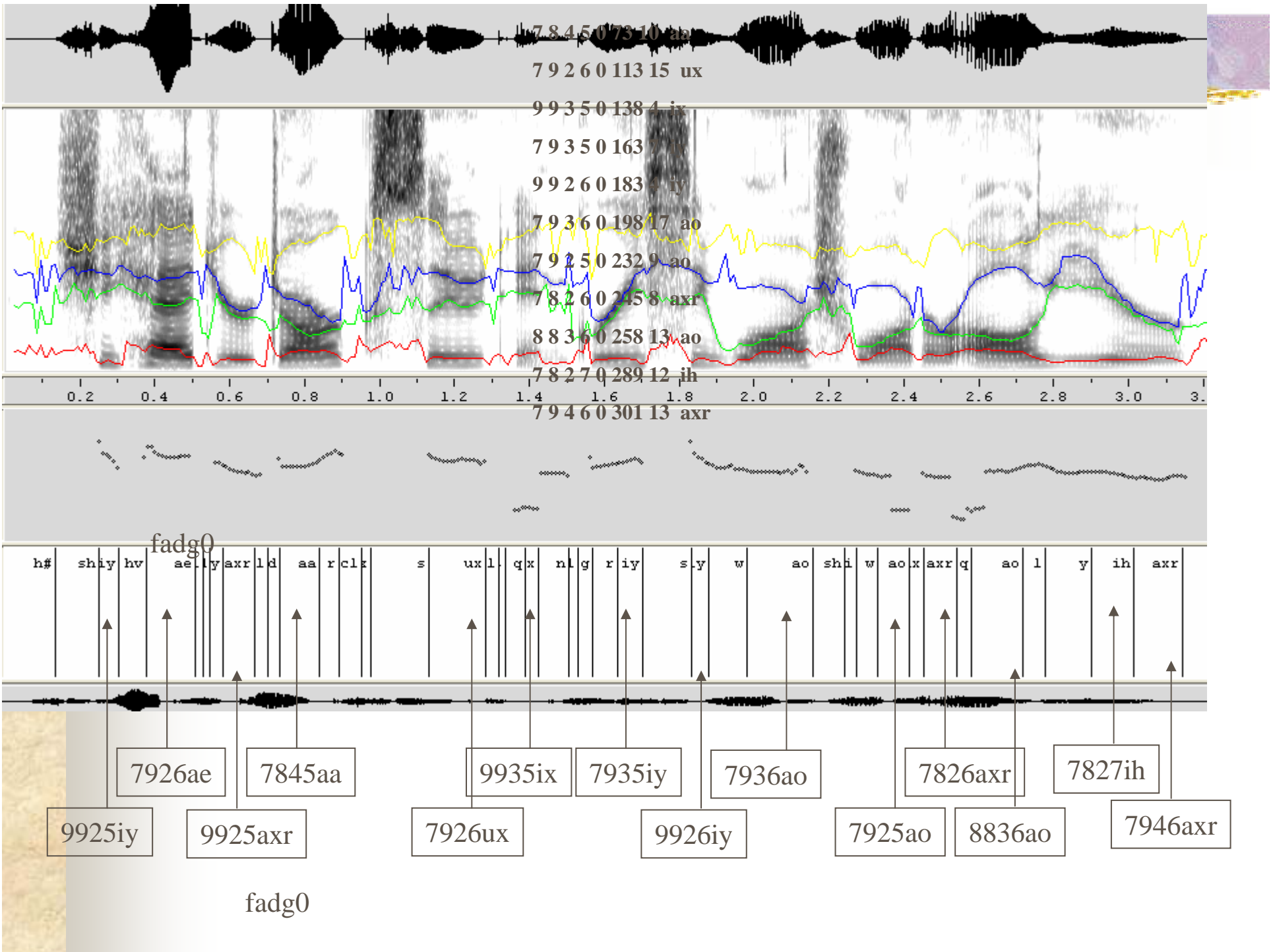


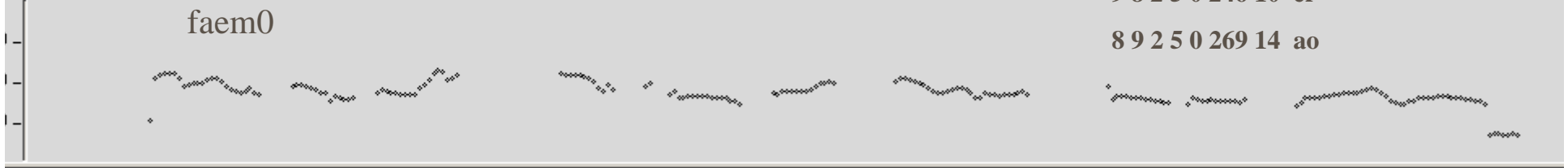
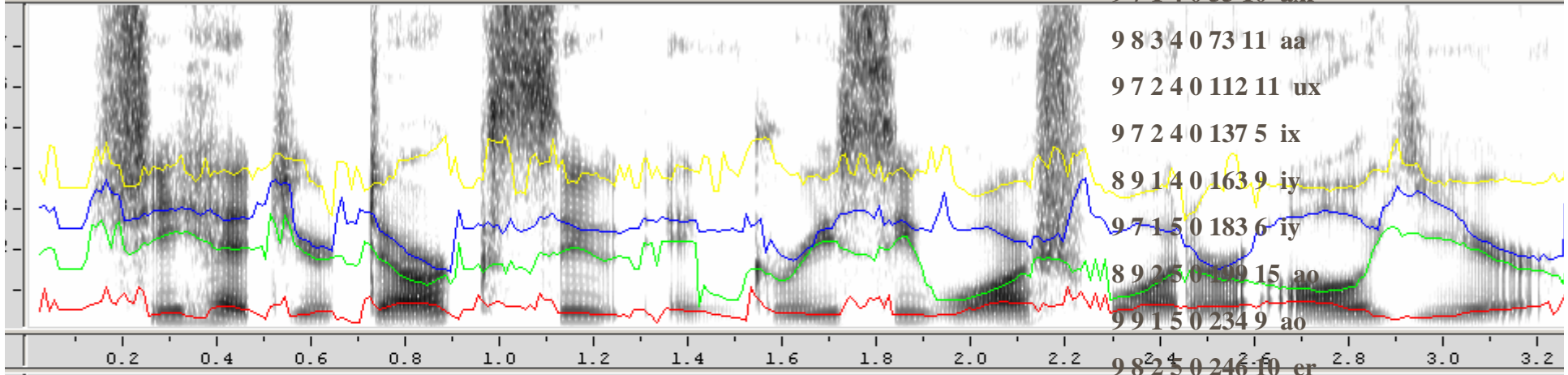
Experiment Result:

The training and testing are both performed on 346 speakers (173 female and 173 male) on the HTIMIT database. 38 MFCC parameters are computed with window size of 30 ms and frame rate of 100 Hz. Nine handsets (cb1-cb4, e11-e14, and pt1) and one Sennheizer head-mounted microphone (senh), from HTIMIT were used as the training and testing handsets in the experiments, respectively the utterance of these handsets were divided into train and test categories by randomly choosing 400 sec utterance as testing data and the other 1600 sec utterance as training data.



	mismatch	mismatch+EPA	mismatch+pitch	mismatch+pitch+EPA
sen	77.16	80.92	77.74	81.05
cb1	71.38	72.96	76.01	77.01
cb2	67.05	69.78	73.69	75.27
cb3	30.05	34.52	26.3	31.34
cb4	41.32	45.5	39.59	45.21
e11	66.18	69.49	72.83	75.56
e12	61.56	62.84	64.16	66.02
e13	59.53	60.82	65.31	68.63
e14	64.45	67.18	70.23	74.12
pt1	52.31	55.91	61.56	64.29
average	57.09	59.89	61.08	64.16
error rate reduction		2.80		3.09

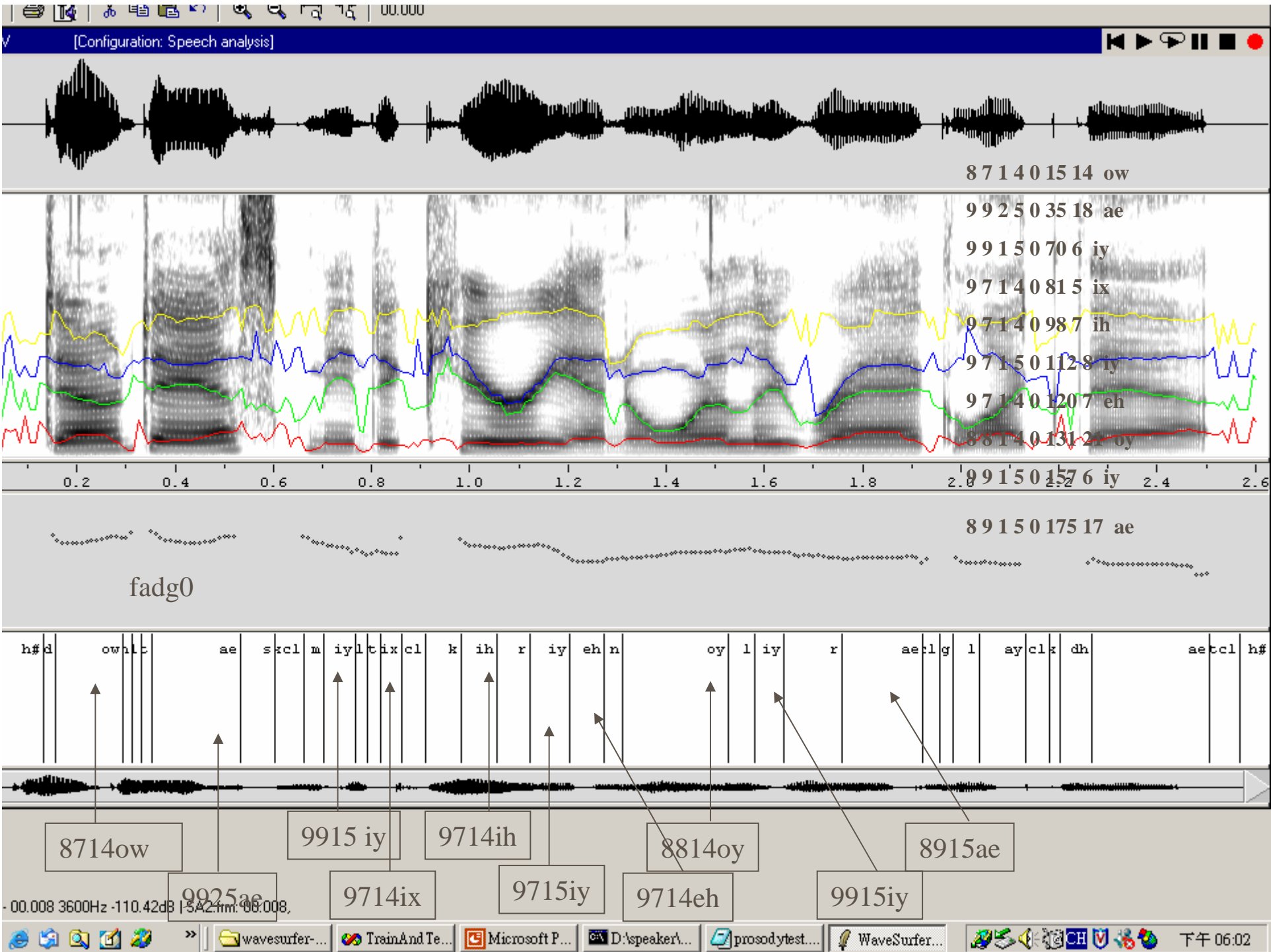


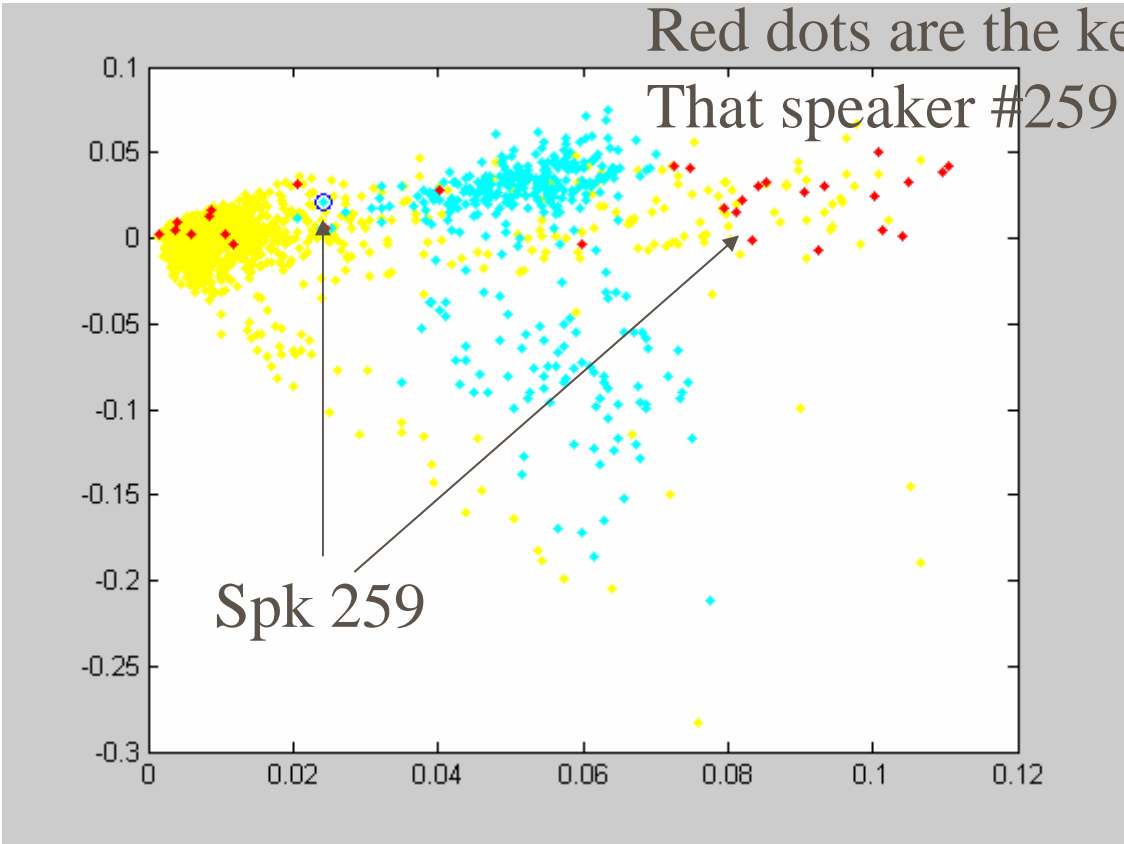


h# sh iy hv eh l h axr dcl aa r:cl: s ux qix ncl gr iy siy w ao shpi w aox er q ao l y ih er



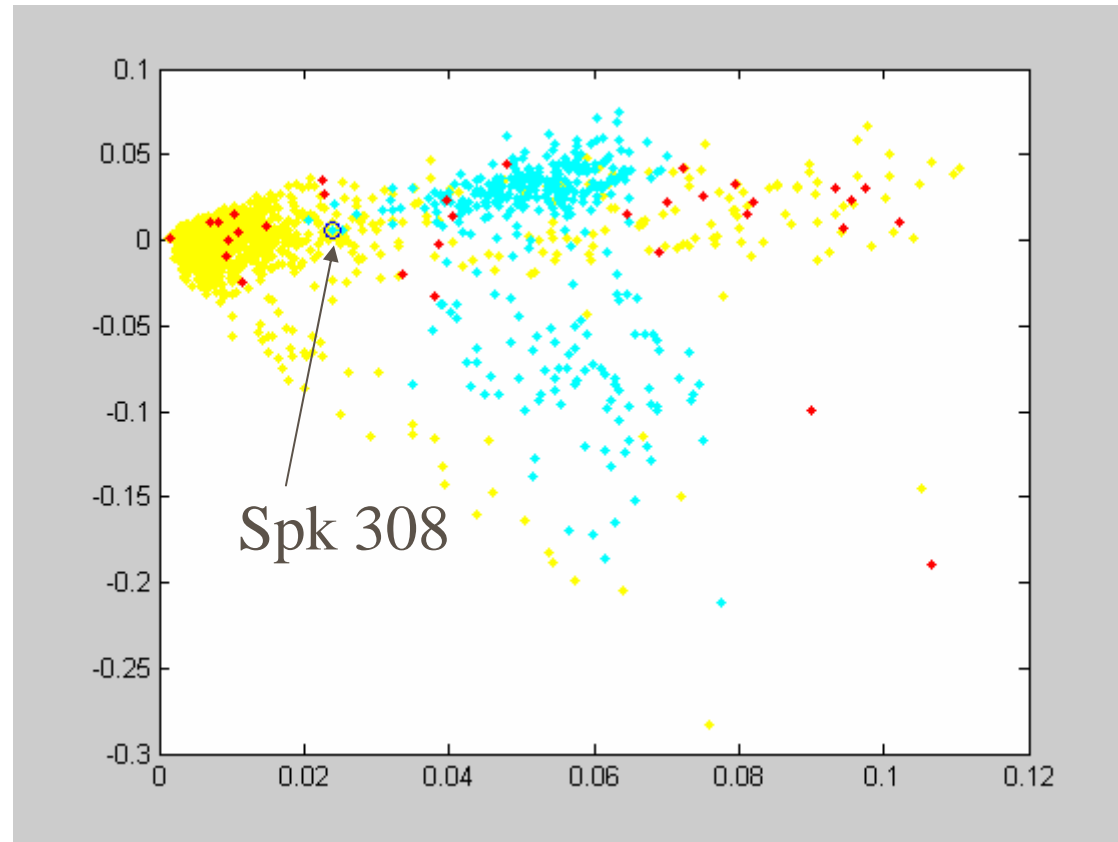
8815iy	9714eh	9834aa	9724ux	8914iy	8925ao	8925er
9714axr	9724ix	9715iy	9915ao	8925ao		

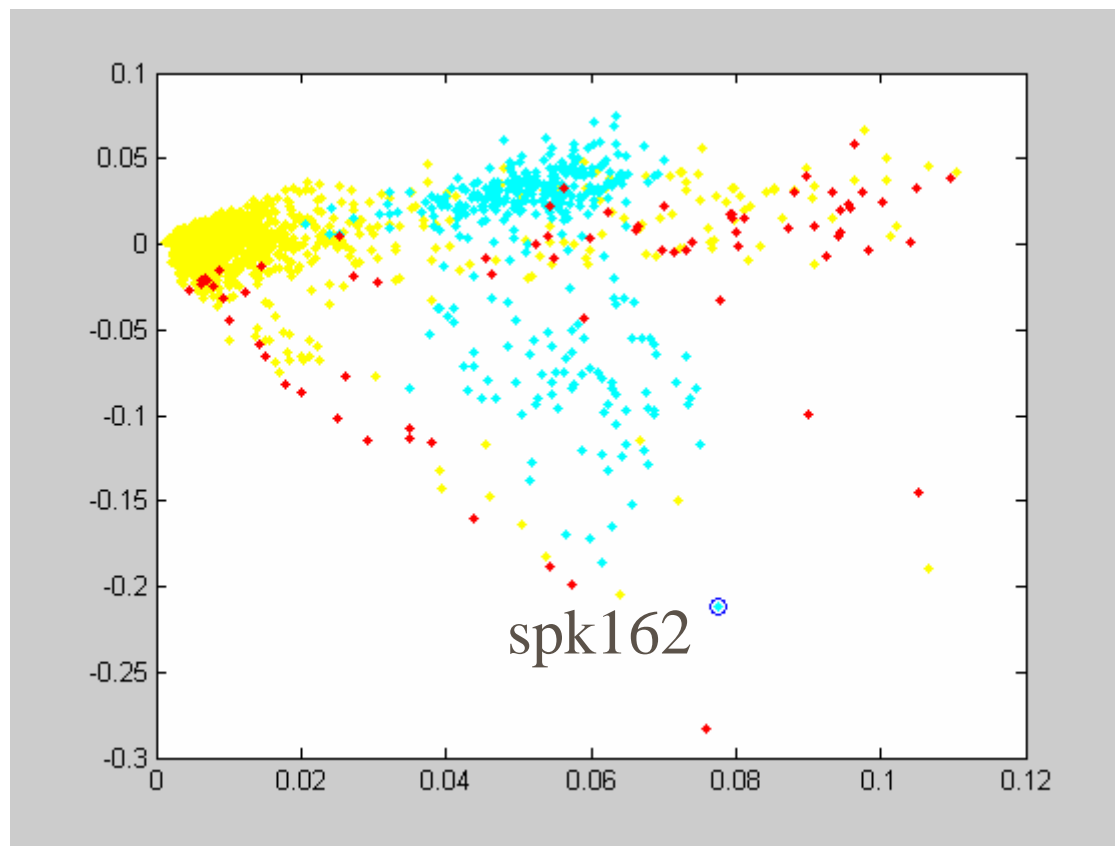


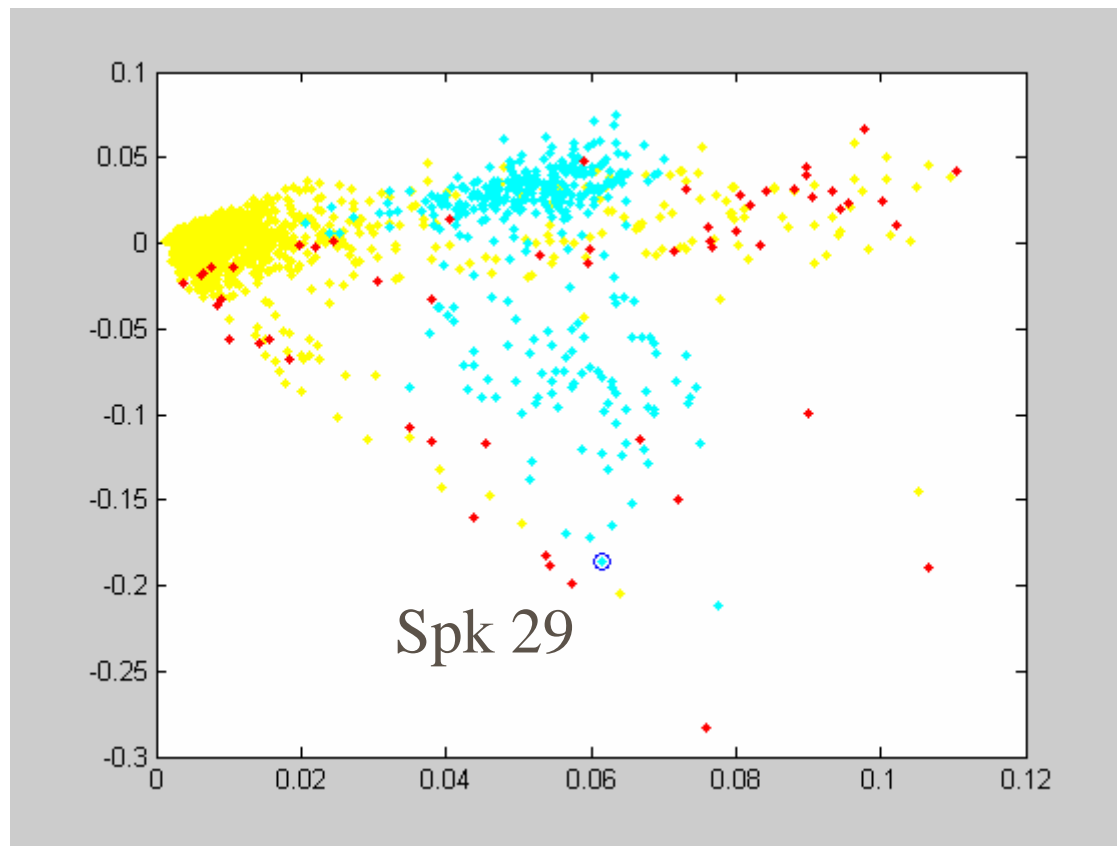


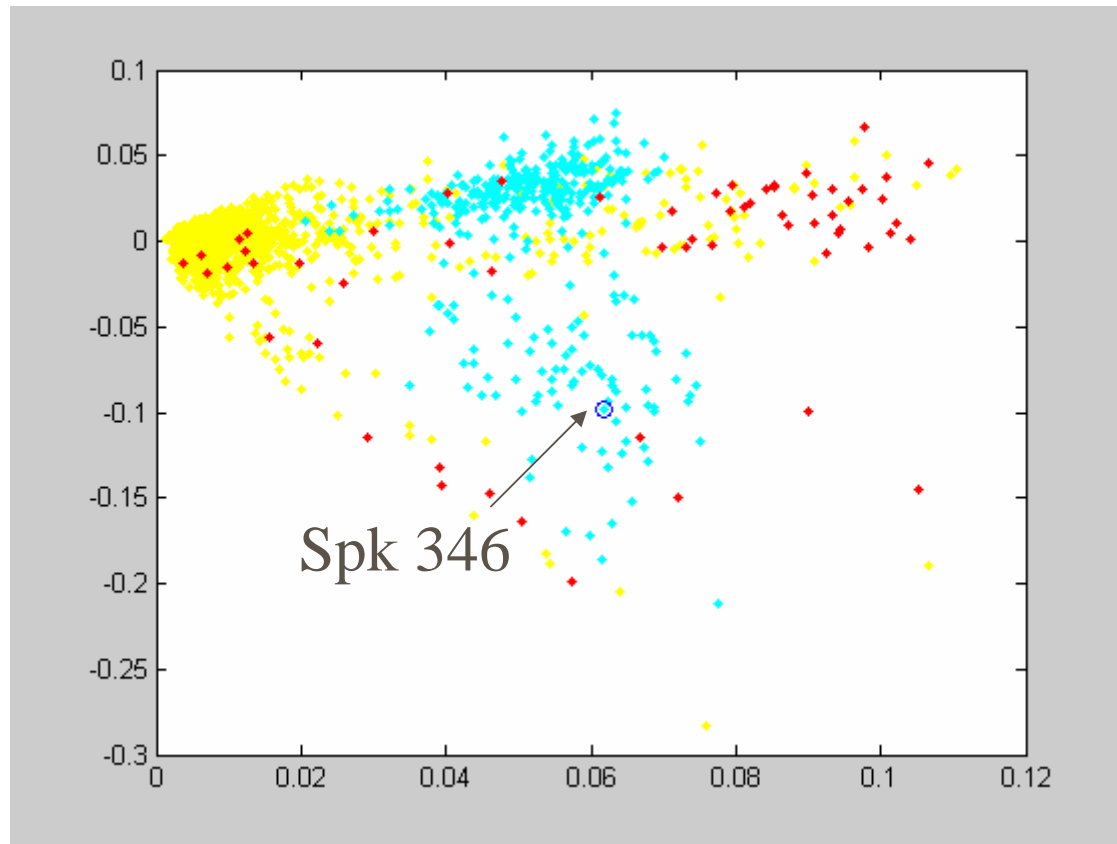
Red dots are the keywords
That speaker #259 has been said

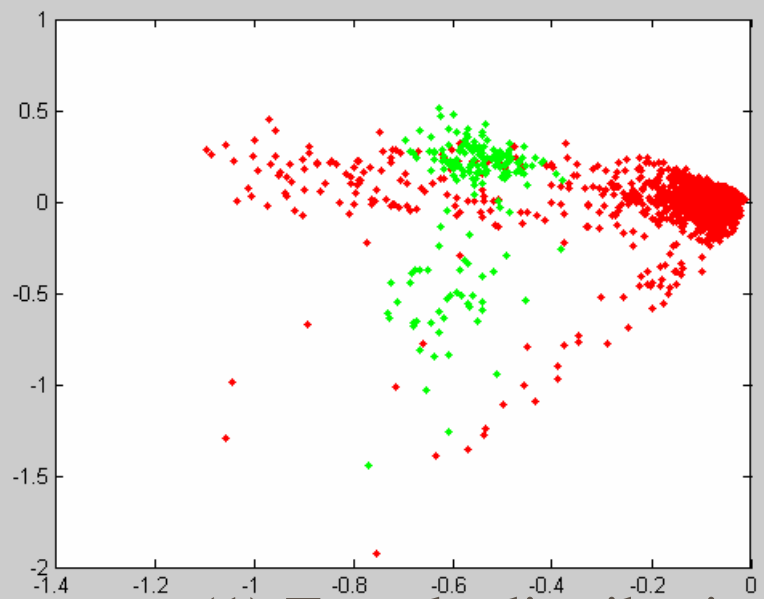
Spk 259



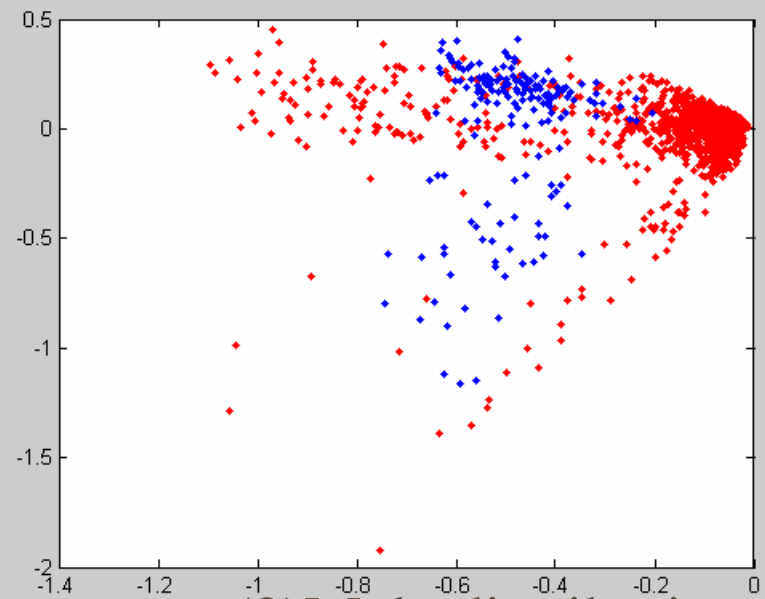




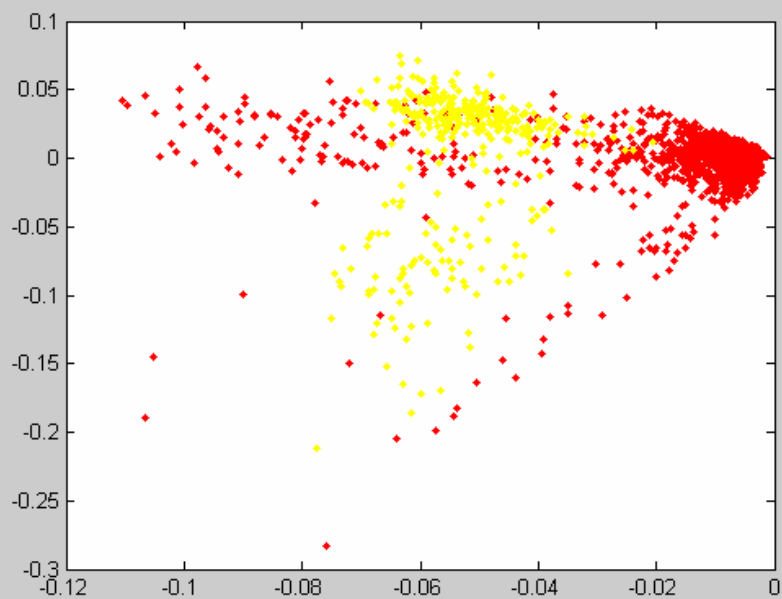




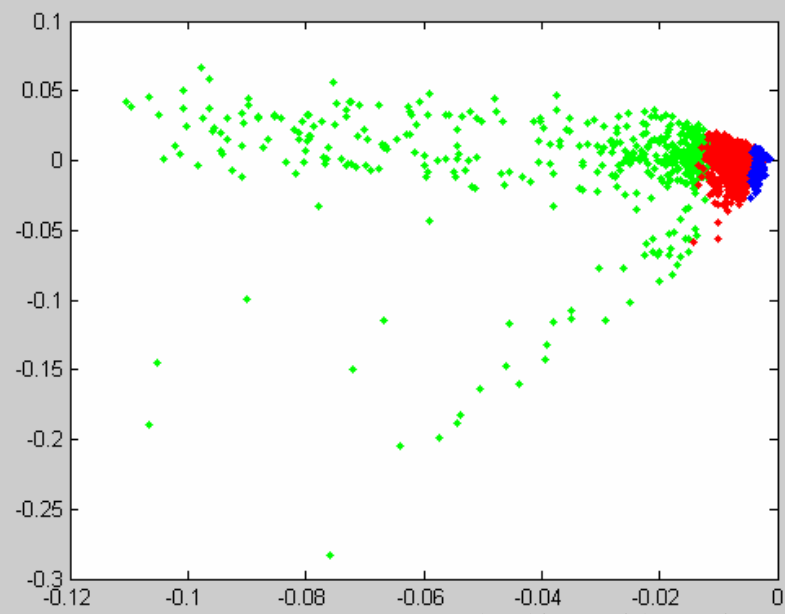
(1) Female distribution



(2) Male distribution



(3) All speaker distribution

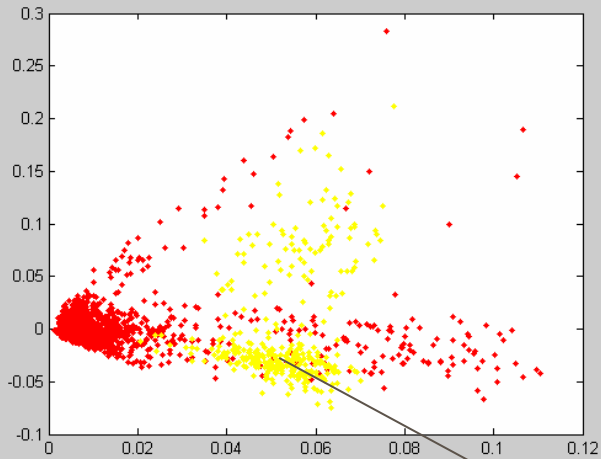


(4) Keyword distribution



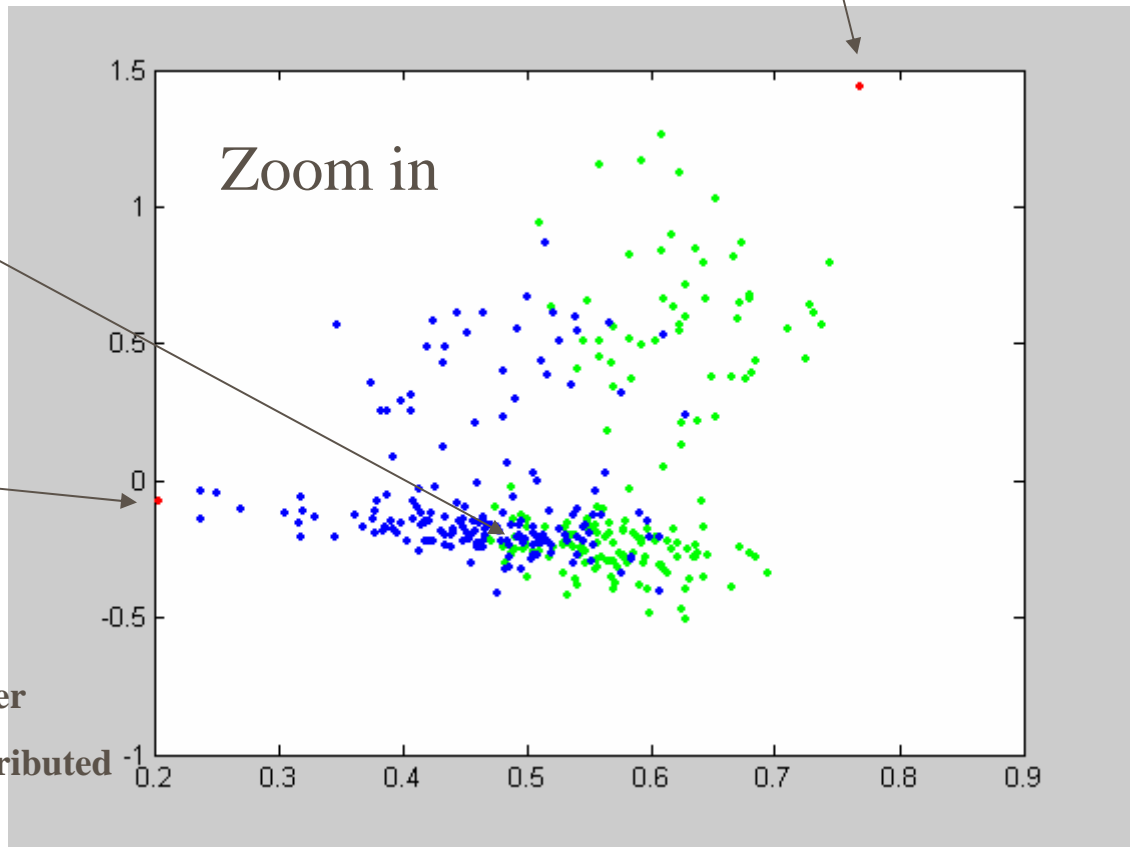
The red dots are represented to all the 3899 keyword

Rank 1

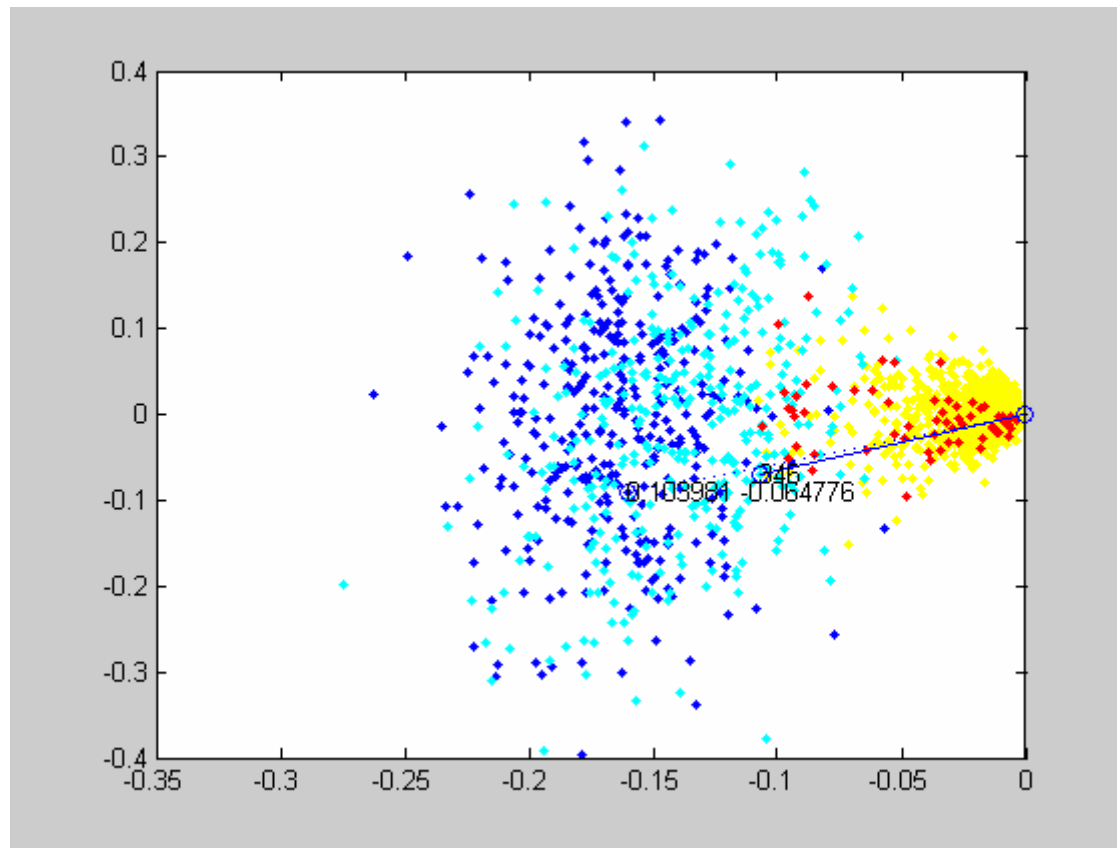


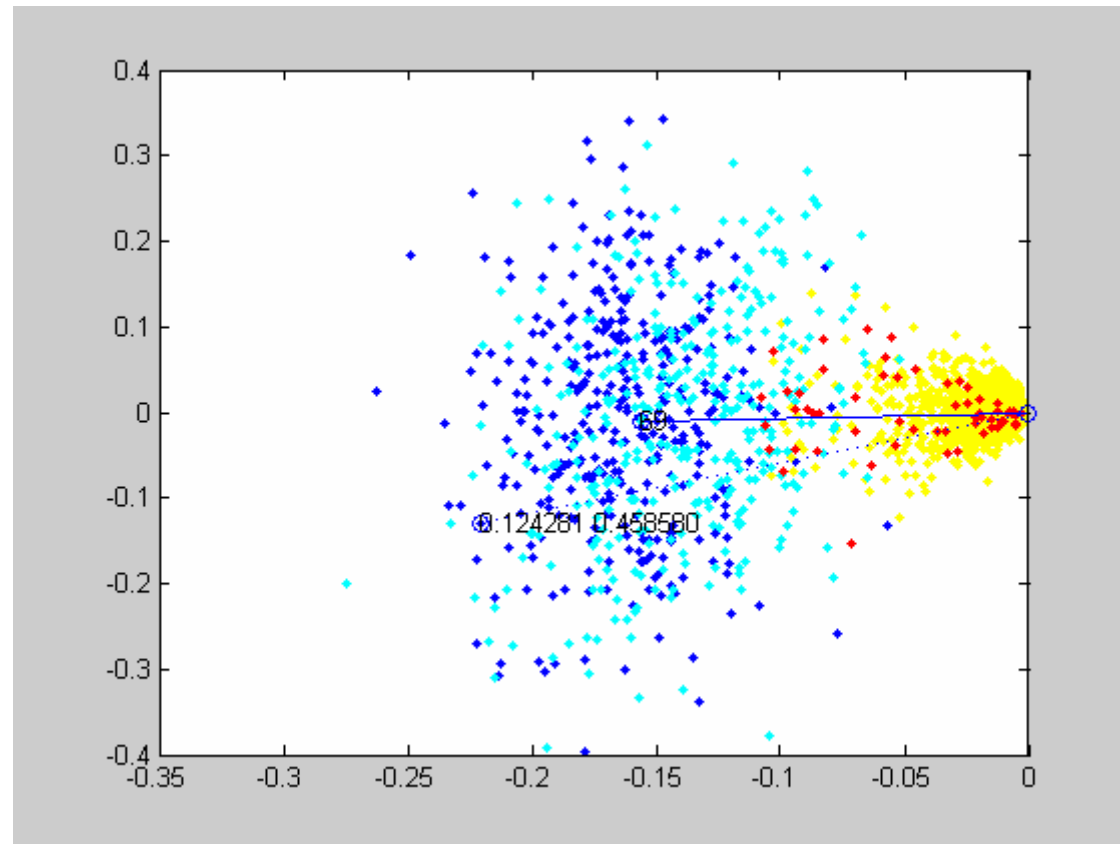
The yellow dots are represented to all the 346 speaker

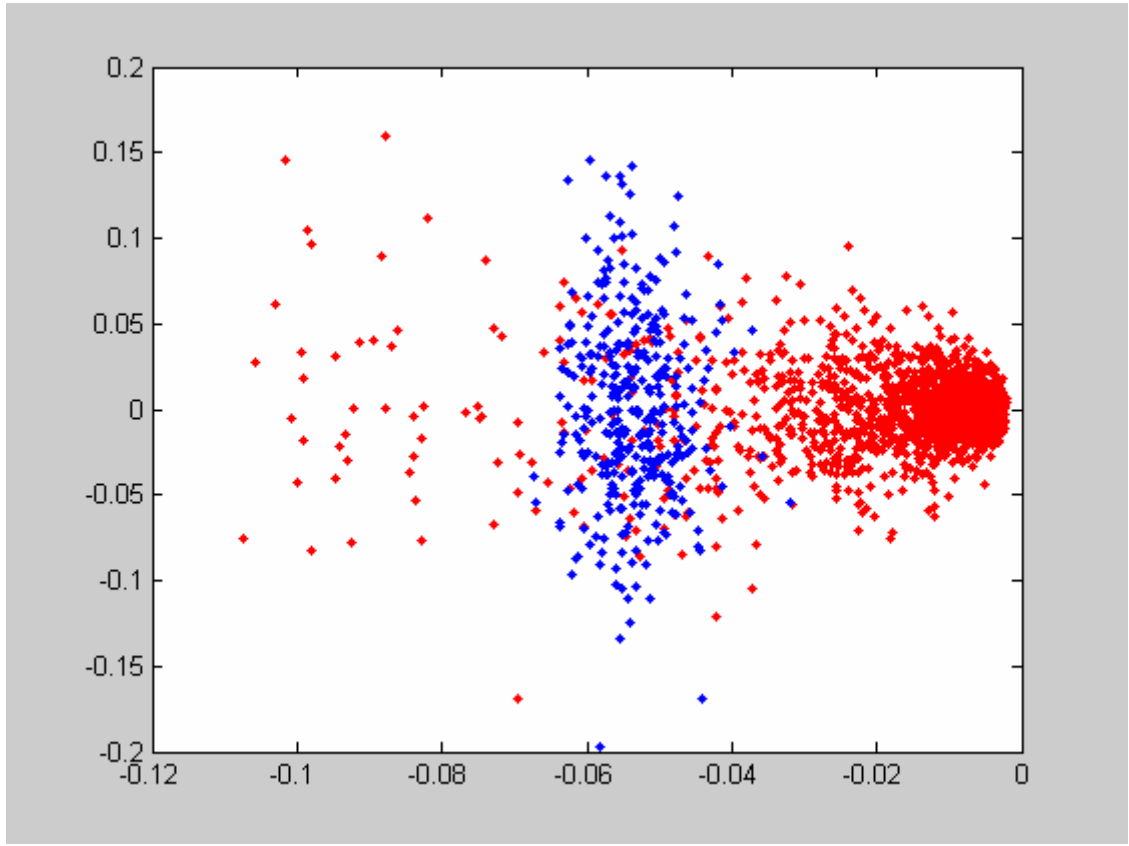
Rank 346

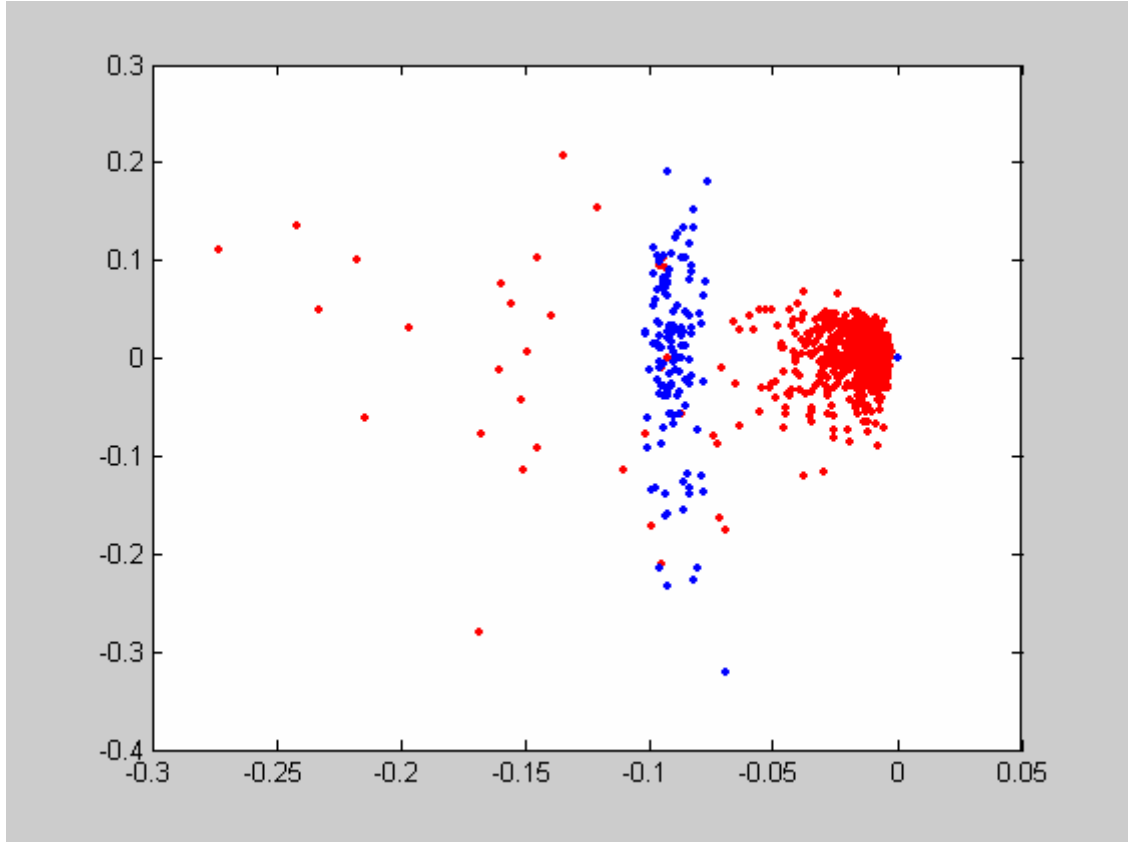


When the V matrix is sorted by incremental order, we found that the speakers in low rank are distributed in blue region and the higher rank speakers are distributed in green region. Moreover, the red dots is the Rank 1 speaker and rank 346 speaker.











Conclusions:

In this research, we address a novel model to extract the prosodic patterns. Then use these patterns to apply the EPA to score the prosodic information of each speaker, which means the long-term information can be retrieve from the prosodic pattern. The traditional GMM score and EPA score has been fusion into a framework to increase the speaker recognition. Especially in mismatch channel condition, it is found that the performance of EPA would affect the entire performance significantly. The possible strategies for improvement in our future work include using as much information as possible, such as pitch jump, and pitch histogram to construct the tree apply to EPA in the future.