Combining Spectral Representations for Large-Vocabulary Continuous Speech Recognition

Author Giulia Garau and Steve Renals

Professor:陳嘉平

Reporter:葉佳璋

Outline

- Introduction
- Pitch-Synchronous Analysis
- STRAIGHT-Based Features
- Feature combination
- Experiments

Introduction

- Different acoustic representations have different strength, and thus will tend to result in ASR systems that make different errors.
- We investigate the combination of complementary acoustic feature streams in large-vocabulary continuous recognition(LVCSR).

Pitch-Synchronous Analysis

- The short time Fourier transform involves the computation of a separate Fourier transform for each frame of the signal waveform under a sliding window.
  - ➢ Long window in time: good resolution in frequency and poorer time resolution
  - ➢ Short window in time: good time resolution and poorer frequency resolution
  - ➢ Fixed size window: its effect will be evident on the spectrum, particularly in high pitch speakers.

Pitch-Synchronous Analysis

- Not broad enough to remove the harmonic structures for high pitched speakers, usually for females.

- Interest to investigate the use of a pitch-synchronous window that adapts according to the use current estimate of the fundamental frequency.

STRAIGHT-Based Features

- STRAIGHT is a vocoder consisting of analysis and synthesis parts.
    - The spectral analysis of STRAIGHT uses a pitch-adaptive window which gives equivalent resolution both in time and frequency domains.
    - An interpolation is then performed on the partial information given by the adaptive windowing.
- Result in a smoothed time-frequency representation which is not affected by interference arising from signal periodicity.

STRAIGHT-Based Features

- We derived STRAIGHT-based MFCCs by replacing the classic STFT.
- STRAIGHT spectral analysis using a window is Gaussian both in time and frequency.

$$\omega(t) = \frac{1}{\tau_0} \exp\left(-\pi (t/\tau_0)^2\right)$$

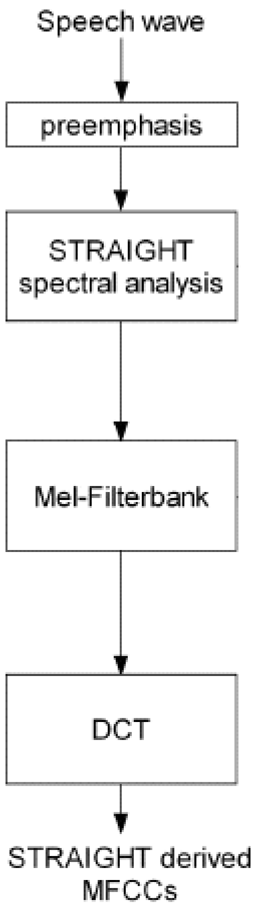
$$W(t) = \frac{\tau_0}{\sqrt{2\pi}} \exp\left(-\pi (\omega/\omega_0)^2\right)$$

- The shape of window depend on the fundamental frequency $f_0 = 1/\tau_0 = 2\pi/\omega_0$

Pitch Tracker

- The value of f_0 used for the window computation can be estimated using various algorithms.
  - TEMPO: obtained using a wavelet Gabor filter designed to highlight the fundamental frequency and reject harmonic replicas
  - PART: based on cross correlation in the time domain

STRAIGHT-derived MFCCs



- Show a block diagram of extraction procedure for STRAIGHT-derived MFCCs.
- MF-PLPs have also been extracted from the log STRAIGHT spectrogram
  - ➢Mel scaling
  - ➢Equal loudness pre-emphasis
  - ➢Cube root compression
  - ➢Linear predictive cepstral analysis

Feature Combination

- Different acoustic representations have different strengths and weaknesses for ASR.
- The simplest form of direct feature combination involves the concatenation of the acoustic feature vectors.
- This approach has a number of drawbacks including a substantial increase in the dimensionality.

Feature Combination

- Both these problems are addressed through the use of dimension reducing decorrelating transforms such as LDA, HLDA, principal components analysis(PCA).

- PCA estimate a global transform and has been found to be much less well-suited to the task compared with LDA and HLDA.

HLDA

- In HLDA and LDA, each feature vector that is used to derive the transformation is assigned to a class, since one of the goals is to improve the discrimination between the classes used during decoding.
  - Given an n dimensional feature vector x, the goal is to find a linear transformation $\theta^T : R^n \rightarrow R^p$ such as to project x in a p dimension space according to $y = \theta^T x$.
  - The transform is chosen to maximize the between class covariance Σ_b and to minimizing the within class covariance Σ_w.
  - Using the eigenvectors corresponding to the p largest eigenvalues of $\Sigma_b \Sigma_w^{-1}$.
- LDA makes two assumptions:
  - All the classes follow a multivariate Gaussian distribution.
  - They share the same within-class covariance matrix.

HLDA

- In HLDA, optimal transformation matrix A is found by maximizing the likelihood of the original data x

$$\log L\left(x;A\right) = -\frac{nN}{2} + \sum_{j=1}^{J} \frac{N_j}{2} \log\left(\frac{\left(\det A\right)^2}{\left(2\pi\right)^n \Pi_{k=1}^{P} a_k \hat{\Sigma}^{(j)} a_k^T \Pi_{k=p+1}^{P} a_k \hat{\Sigma}^{(j)} a_k^T} \right)$$

  - $\hat{\Sigma}$ and $\hat{\Sigma}^{(j)}$: the global and per class covariance matrix
  - N and N_j : are the number of corresponding training vectors

- In which the transform matrix A is computed by periodically reestimating individual rows a_k

$$\hat{a}_k = c_k G^{(k)-1} \sqrt{\frac{N}{c_k G^{(k)-1} c_k^T}} \qquad G^{(k)} = \begin{cases} \Sigma_{j=1}^{J} \dfrac{r_j}{a_k \hat{\Sigma}^{(j)} a_k^T} \hat{\Sigma}^{(j)}, k \le p \\ \dfrac{N}{a_k \hat{\Sigma} a_k^T} \hat{\Sigma}, k > p \end{cases}$$

  - c_i : ith row vector of matrix $C = |A| A^{-1}$ for the current estimate of A
  - r_j :the number of training feature vectors belonging to the jth class

HLDA

- In order to avoid data sparsity, the type of classes used to estimate the HLDA transformation matrix should be carefully considered.
- We experimented with two possible choice of classes
    - ➢ Classes corresponding to the HMM triphone states of our models.
    - ➢ Gaussian mixture components of monophone models.

System-level combination

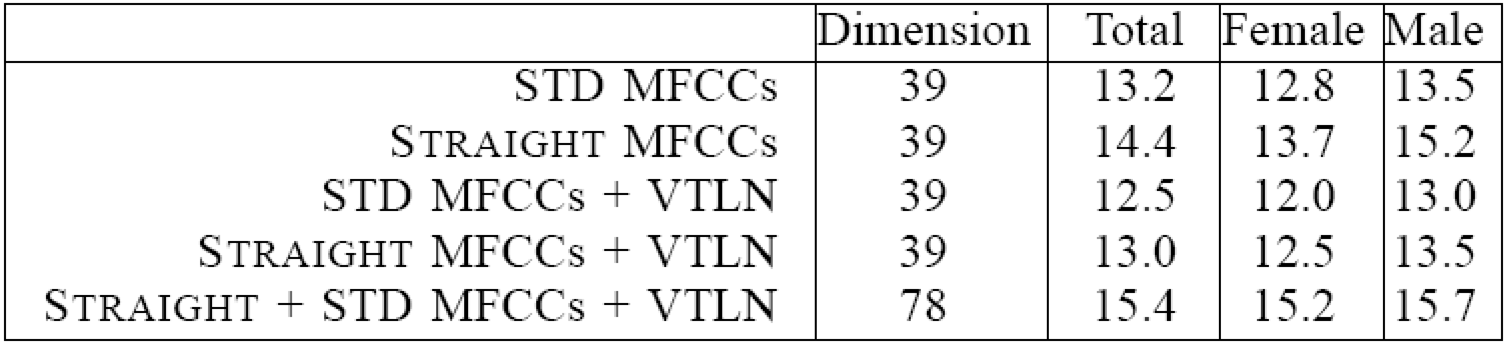
- We also explored the use of system-level combination using ROVER.
  - First compared by aligning them using dynamic programming to minimizing the number of substitutions, deletions, and insertions.
  - Alignments are then combined using a voting approach
    - Selecting the most frequently recognized hypothesis
    - Selecting the highest confident score

Experiments

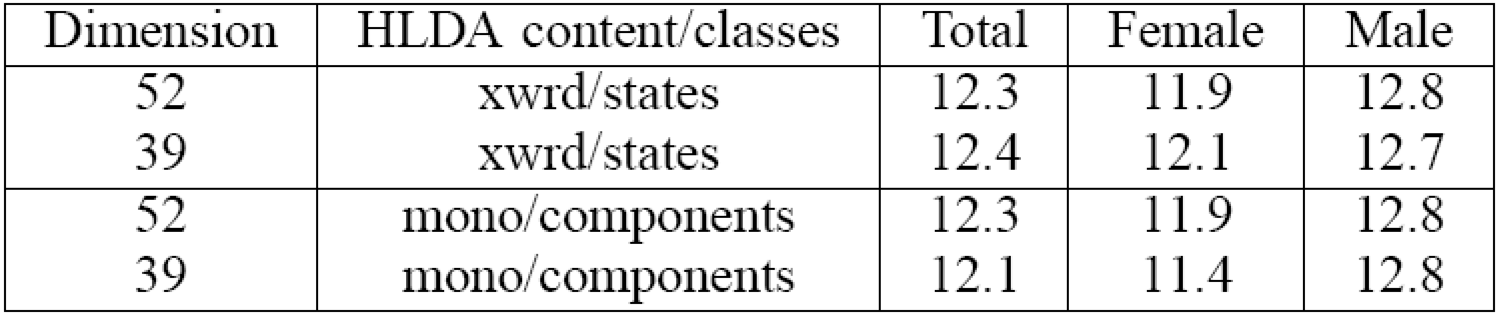
- The baseline acoustic models were trained on conventional MFCC.
  - 25-ms window with 10-ms shift
  - 12 ceptstral coefficients plus the zeroth ceptral coefficient(C0) were estimate, and first and second derivatives were also computed resulting in a 39-element feature vector.
- VTLN was applied both the MFCC and to the STRAIGHT-derived MFCC system.

WSJCAM0

- Our first set of experiments was performed on the WSJCAM0 corpus.
    - ➢Recorded at Cambridge University, and consisting of native British English read speech, using text from the Wall Street Journal corpus.
    - ➢Training set consisting of 7861 utterances, corresponding to around 15 h of speech.
    - ➢We test on the 20000 words “open vocabulary”.



	Dimension	Total	Female	Male
STD MFCCs	39	13.2	12.8	13.5
Straight MFCCs	39	14.4	13.7	15.2
STD MFCCs + VTLN	39	12.5	12.0	13.0
Straight MFCCs + VTLN	39	13.0	12.5	13.5
Straight + STD MFCCs + VTLN	78	15.4	15.2	15.7

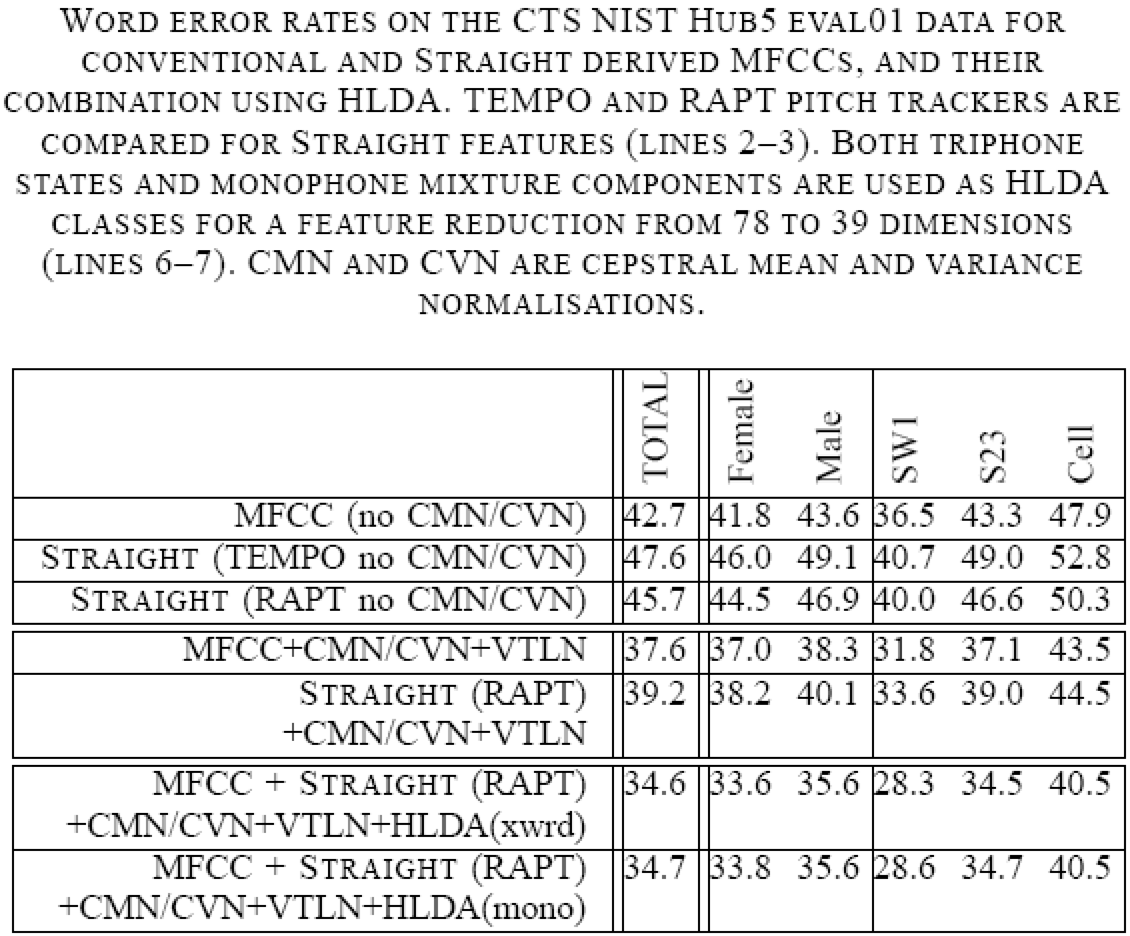


Dimension	HLDA content/classes	Total	Female	Male
52	xwrd/states	12.3	11.9	12.8
39	xwrd/states	12.4	12.1	12.7
52	mono/components	12.3	11.9	12.8
39	mono/components	12.1	11.4	12.8

- The xwrd/states shows the results obatained when HLDA statistics were estimate using states of the cross-word triphone HMMs.
- The mono/components shows the results obtained using monophone mixture components as classes

Conventional Telephone Speech

- CTS data, based on a 72-h training set containing 57h Switchboard-1(SW1), 8h from Switchboard-2(S23), and 7h from Call Home English corpus(Cell).
- Our test set was the NIST Hub5 Eval101 evaluation set consisting of around 6 h of speech in total.



Word error rates on the CTS NIST Hub5 eval01 data for conventional and Straight derived MFCCs, and their combination using HLDA. TEMPO and RAPT pitch trackers are compared for Straight features (lines 2–3). Both triphone states and monophone mixture components are used as HLDA classes for a feature reduction from 78 to 39 dimensions (lines 6–7). CMN and CVN are cepstral mean and variance normalisations.

	TOTAL	Female	Male	SW1	S23	Cell
MFCC (no CMN/CVN)	42.7	41.8	43.6	36.5	43.3	47.9
Straight (TEMPO no CMN/CVN)	47.6	46.0	49.1	40.7	49.0	52.8
Straight (RAPT no CMN/CVN)	45.7	44.5	46.9	40.0	46.6	50.3
MFCC+CMN/CVN+VTLN	37.6	37.0	38.3	31.8	37.1	43.5
Straight (RAPT) +CMN/CVN+VTLN	39.2	38.2	40.1	33.6	39.0	44.5
MFCC + Straight (RAPT) +CMN/CVN+VTLN+HLDA(xwrd)	34.6	33.6	35.6	28.3	34.5	40.5
MFCC + Straight (RAPT) +CMN/CVN+VTLN+HLDA(mono)	34.7	33.8	35.6	28.6	34.7	40.5

Multiparty Meeting

- Four corpora of multiparty meeting recordings:
  - ICSI: 70 h
  - NIST: 13 h
  - CMU-ISL: 16 h
  - AMI: 16 h
- Two principal testing condition :
  - Individual headset microphone(IHM)
  - Multiple distance microphone (MDM) included Wiener filtering

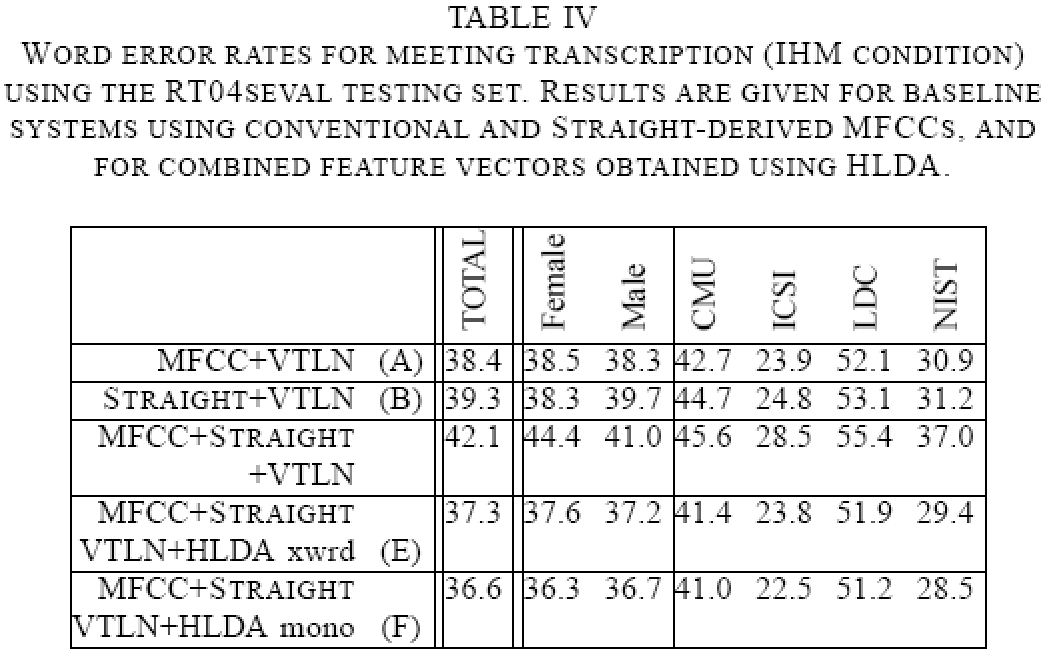
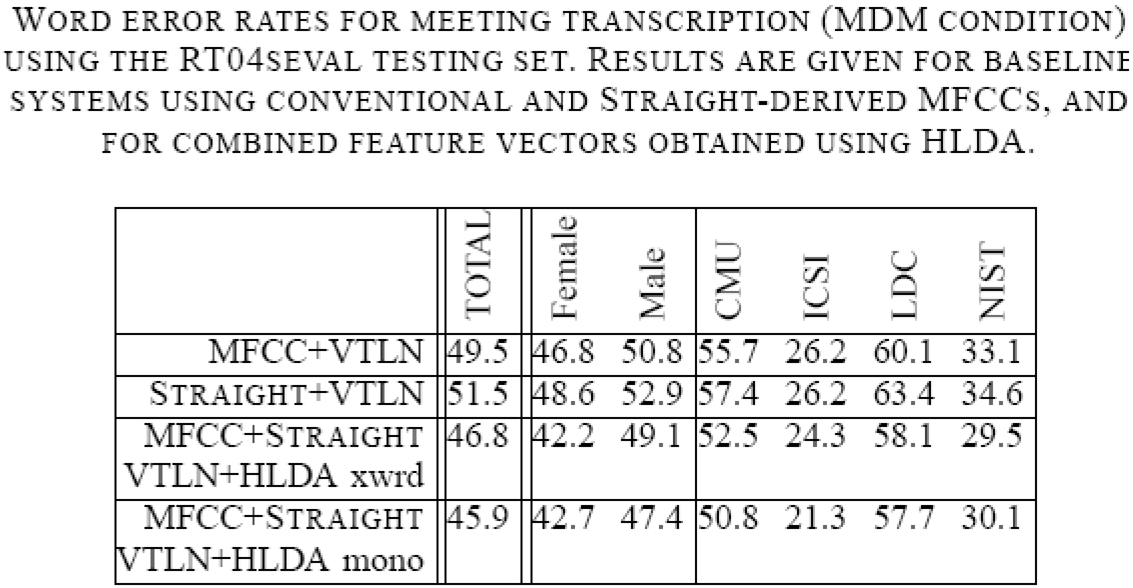


TABLE IV

Word error rates for meeting transcription (IHM condition) using the RT04seval testing set. Results are given for baseline systems using conventional and Straight-derived MFCCs, and for combined feature vectors obtained using HLDA.

	TOTAL	Female	Male	CMU	ICSI	LDC	NIST
MFCC+VTLN (A)	38.4	38.5	38.3	42.7	23.9	52.1	30.9
Straight+VTLN (B)	39.3	38.3	39.7	44.7	24.8	53.1	31.2
MFCC+Straight +VTLN	42.1	44.4	41.0	45.6	28.5	55.4	37.0
MFCC+Straight VTLN+HLDA xwrd (E)	37.3	37.6	37.2	41.4	23.8	51.9	29.4
MFCC+Straight VTLN+HLDA mono (F)	36.6	36.3	36.7	41.0	22.5	51.2	28.5

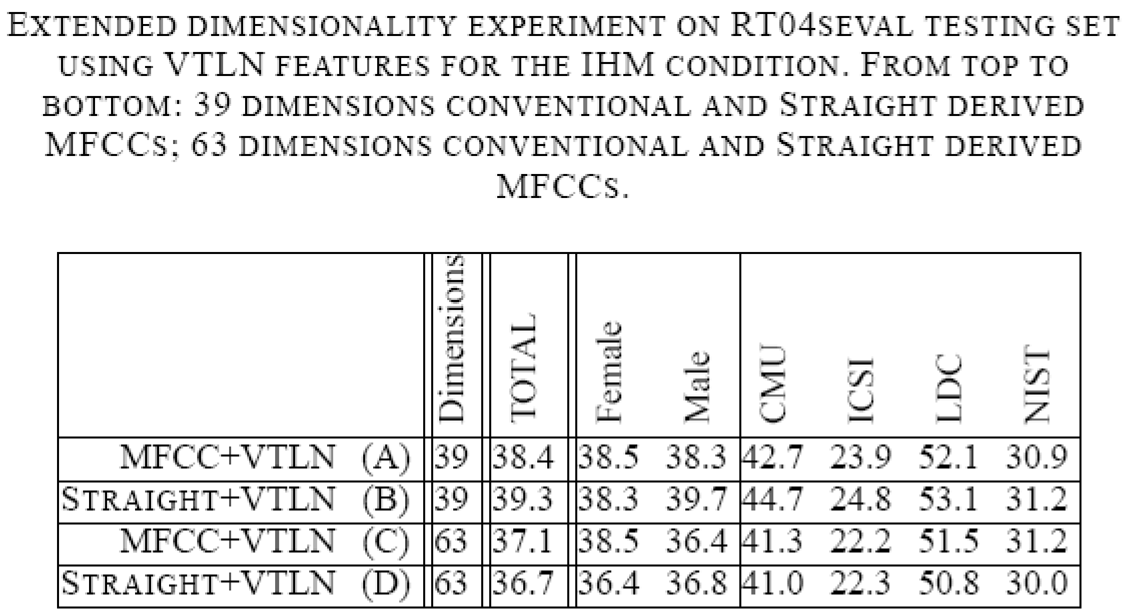


Word error rates for meeting transcription (MDM condition) using the RT04seval testing set. Results are given for baseline systems using conventional and Straight-derived MFCCs, and for combined feature vectors obtained using HLDA.

	TOTAL	Female	Male	CMU	ICSI	LDC	NIST
MFCC+VTLN	49.5	46.8	50.8	55.7	26.2	60.1	33.1
Straight+VTLN	51.5	48.6	52.9	57.4	26.2	63.4	34.6
MFCC+Straight VTLN+HLDA xwrd	46.8	42.2	49.1	52.5	24.3	58.1	29.5
MFCC+Straight VTLN+HLDA mono	45.9	42.7	47.4	50.8	21.3	57.7	30.1

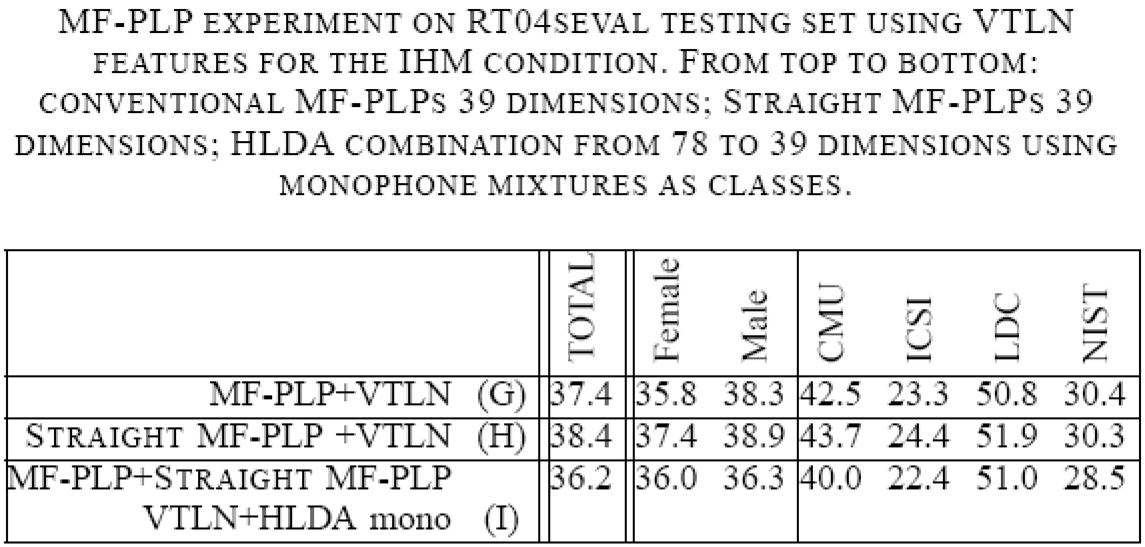
Further Experiment on Meetings

- Higher order cepstral coefficient are known to be the most affected by the spectral harmonic component due to the pitch.
- The smoothed STRAIGHT spectral, which is not affected by spectral harmonic
  - 20 ceptral coefficients plus C0 and their first and second temporal derivatives, result in 63 dimensions
  - IHM meeting domain
- We also performed experiment on the use of STRAIGHT for MF-PLP extraction and ROVER.



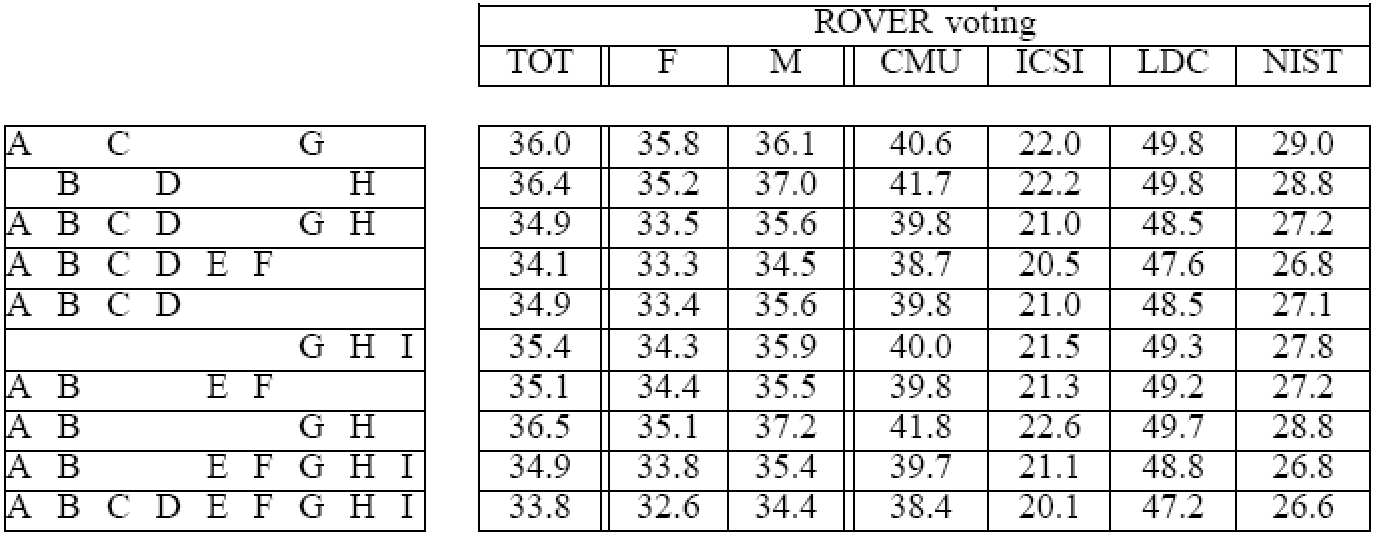
Extended dimensionality experiment on RT04seval testing set using VTLN features for the IHM condition. From top to bottom: 39 dimensions conventional and Straight derived MFCCs; 63 dimensions conventional and Straight derived MFCCs.

	Dimensions	TOTAL	Female	Male	CMU	ICSI	LDC	NIST
MFCC+VTLN (A)	39	38.4	38.5	38.3	42.7	23.9	52.1	30.9
Straight+VTLN (B)	39	39.3	38.3	39.7	44.7	24.8	53.1	31.2
MFCC+VTLN (C)	63	37.1	38.5	36.4	41.3	22.2	51.5	31.2
Straight+VTLN (D)	63	36.7	36.4	36.8	41.0	22.3	50.8	30.0

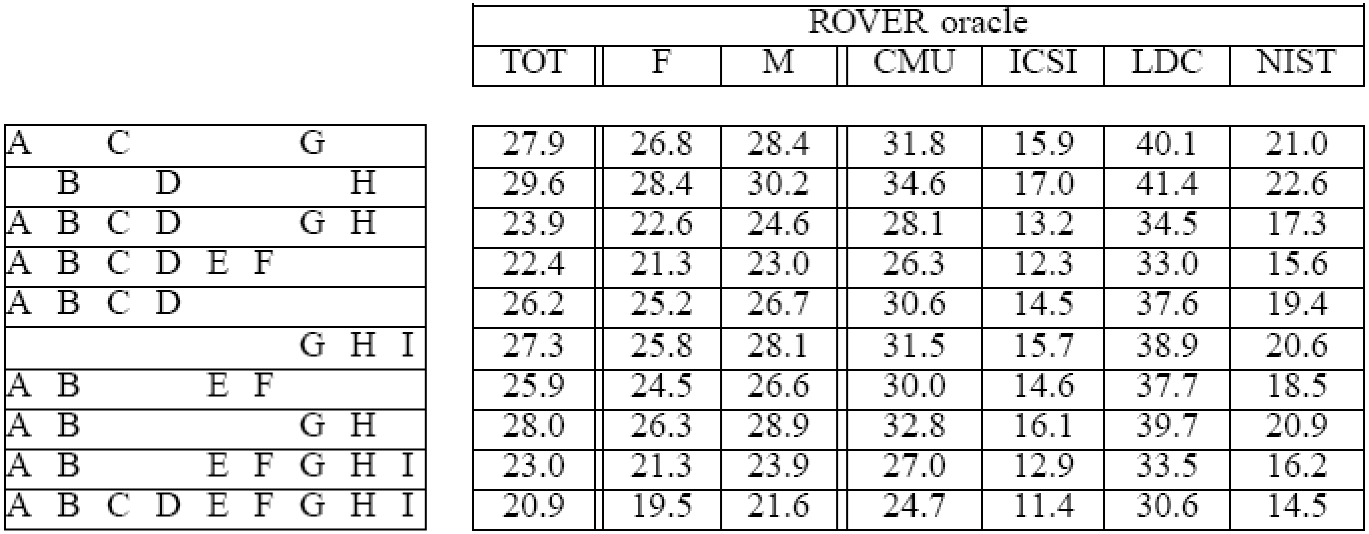


MF-PLP experiment on RT04seval testing set using VTLN features for the IHM condition. From top to bottom: conventional MF-PLPs 39 dimensions; Straight MF-PLPs 39 dimensions; HLDA combination from 78 to 39 dimensions using monophone mixtures as classes.

	TOTAL	Female	Male	CMU	ICSI	LDC	NIST
MF-PLP+VTLN (G)	37.4	35.8	38.3	42.5	23.3	50.8	30.4
Straight MF-PLP +VTLN (H)	38.4	37.4	38.9	43.7	24.4	51.9	30.3
MF-PLP+Straight MF-PLP VTLN+HLDA mono (I)	36.2	36.0	36.3	40.0	22.4	51.0	28.5



	ROVER voting						
	TOT	F	M	CMU	ICSI	LDC	NIST
A C G	36.0	35.8	36.1	40.6	22.0	49.8	29.0
B D H	36.4	35.2	37.0	41.7	22.2	49.8	28.8
A B C D G H	34.9	33.5	35.6	39.8	21.0	48.5	27.2
A B C D E F	34.1	33.3	34.5	38.7	20.5	47.6	26.8
A B C D	34.9	33.4	35.6	39.8	21.0	48.5	27.1
G H I	35.4	34.3	35.9	40.0	21.5	49.3	27.8
A B E F	35.1	34.4	35.5	39.8	21.3	49.2	27.2
A B G H	36.5	35.1	37.2	41.8	22.6	49.7	28.8
A B E F G H I	34.9	33.8	35.4	39.7	21.1	48.8	26.8
A B C D E F G H I	33.8	32.6	34.4	38.4	20.1	47.2	26.6



	ROVER oracle						
	TOT	F	M	CMU	ICSI	LDC	NIST
A C G	27.9	26.8	28.4	31.8	15.9	40.1	21.0
B D H	29.6	28.4	30.2	34.6	17.0	41.4	22.6
A B C D G H	23.9	22.6	24.6	28.1	13.2	34.5	17.3
A B C D E F	22.4	21.3	23.0	26.3	12.3	33.0	15.6
A B C D	26.2	25.2	26.7	30.6	14.5	37.6	19.4
G H I	27.3	25.8	28.1	31.5	15.7	38.9	20.6
A B E F	25.9	24.5	26.6	30.0	14.6	37.7	18.5
A B G H	28.0	26.3	28.9	32.8	16.1	39.7	20.9
A B E F G H I	23.0	21.3	23.9	27.0	12.9	33.5	16.2
A B C D E F G H I	20.9	19.5	21.6	24.7	11.4	30.6	14.5