The ISL RT-07 Speech-to-Text System

Matthias Wölfel, Sebastian Stüker, and Florian Kraft

Institut für Theoretische Informatik
Universität Karlsruhe (TH)

Improvements over our RT-06S System

- Automatic Segmentation and Clustering
- Signal Adaptive Front-End
- Channel Selection (no time for joined approach with combination)
- Cross System Adaptation and Combination Varying Both:
  The Front-End and the Phoneme Set (not presented)
- Improved Acoustic Models due to MMIE Training (not presented)
- Experiments and Error Analysis

Automatic Segmentation and Clustering

Automatic segmentation for the various conditions of the lecture subtasks is provided by different systems:

- **Individual head-mounted (IHM)**

  we relied on the segmentation and speaker clusters provided by ICSI (Thanks!)

- **Single distant microphone (SDM) and multi distant microphone (MDM)**

  - *To cut the speech into parts we performed decoder based segmentationused which is a multi-microphone extended version of the this years EPPS transcription system developed and evaluated under the TC-STAR project.*
  - *For segmentation we used the same hierarchical, agglomerative clustering technique as last year which is based on TGMM-GLR distance measurement and the Bayesian information criterion stopping criteria*

Bilinear Transformation
Review

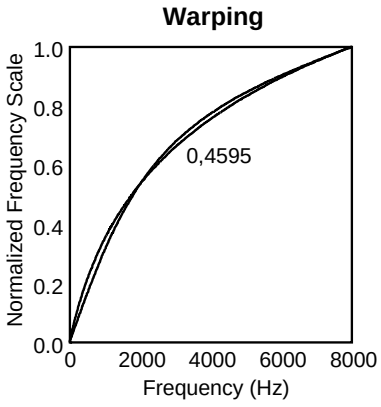
Idea: Replace a unit delay element with a bilinear transformation to warp the frequency axis

$$e^{-j\tilde{\omega}} = D_1(e^{-j\omega}) = \frac{e^{-j\omega} - \alpha}{1 - \alpha \cdot e^{-j\omega}}$$

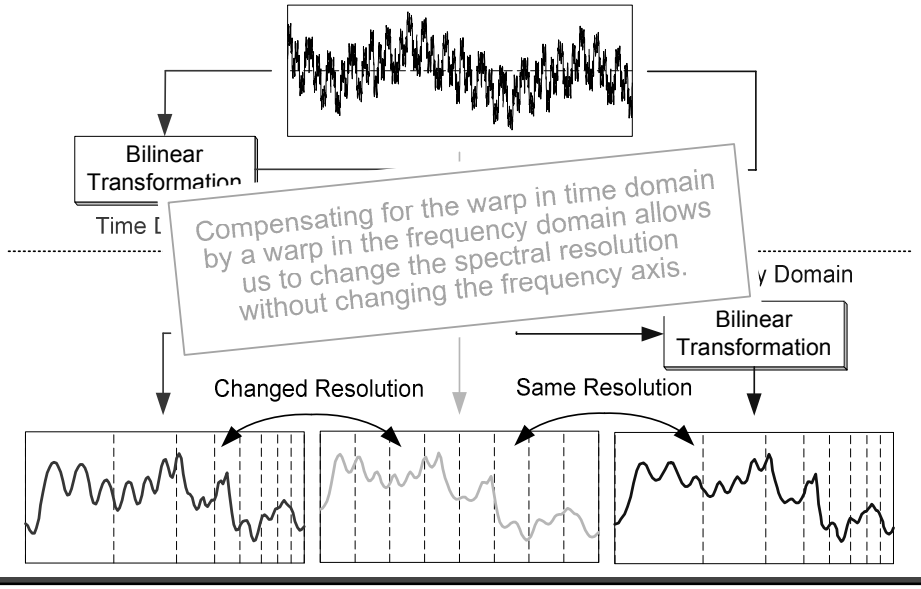
It is possible to approximate the Mel-scale!

Black line: Mel-scale

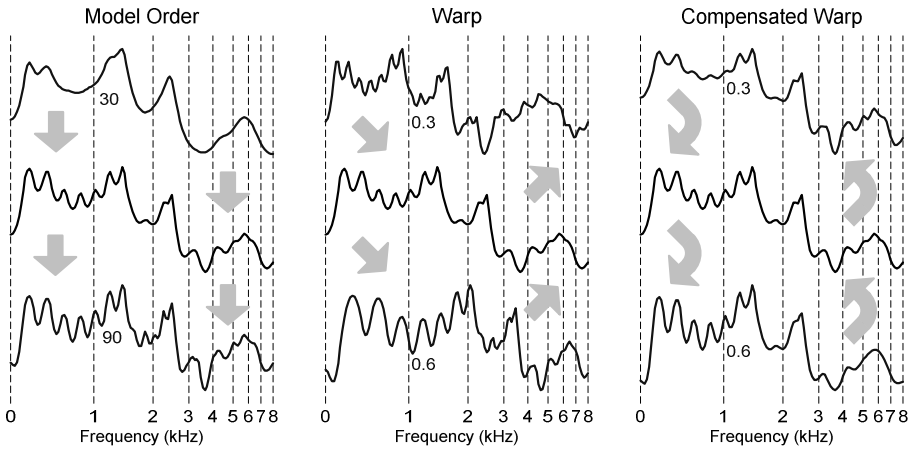
Red line: bilinear transformation with α = 0,4595



Bilinear Transformation
Time vs. Frequency Domain



Influence of the Free Parameters
Model Order, Warp and Compensated Warp



The arrows are showing the influence of the free parameters of the warped-twice MVDR spectral envelope pointing in the direction of higher resolution. The black line is showing an envelope with model order 60 and warp factor 0,4595.

Steering Function

Goal: We wish to adapt our spectral envelope estimate by the free parameters, so that classification relevant characteristics are emphasized while less relevant information is suppressed.

Solution: One promising approach to steer the spectral resolution to lower or higher frequencies was suggested in the work by Nakatoh et al.

$$\alpha_i = +\gamma \cdot \left(\left| \frac{R_i[1]}{R_i[0]} \right| - 0{,}5 \right) + \alpha_{Mel}$$

$R_i[0]$: zero autocorrelation coefficient

$R_i[1]$: first autocorrelation coefficient

γ : sensibility set to 0,1

i : frame index

Class Separability and WER
close talking RT05 develop. and evaluation

Spectrum	Order	Cepstra	Class Separability			Word Error Rate %					
Test Set			Train	Develop	Eval	Develop			Eval		
Pass						1	2	3	1	2	3
Fourier	–	13	11.007	16.470	16.088	36.1	30.3	28.0	35.3	29.7	27.7
Fourier	–	20	11.620	17.929	16.299	36.0	29.7	27.7	37.2	31.3	28.4
WMVDR	60	13	10.768	16.813	16.261	35.0	30.0	28.2	35.5	29.9	27.6
WMVDR	60	20	11.337	18.022	16.614	34.5	29.1	27.3	35.3	29.6	27.3
WMVDR	30	13	10.900	17.675	16.702	34.6	29.8	27.8	34.7	29.6	27.2
WMVDR	30	20	11.386	18.630	17.318	33.9	29.1	27.4	34.9	29.2	26.9
W2MVDR	60	13	10.893	17.673	16.456	34.5	29.5	27.5	34.1	29.2	27.0
W2MVDR	60	20	11.473	18.510	16.818	34.1	28.8	26.8	35.4	29.0	26.3

used in evaluation

Class Separability and WER
distant talking RT05 develop. and evaluation

Spectrum	Order	Cepstra	Class Separability		Word Error Rate %					
Test Set			Develop	Eval	Develop			Eval		
Pass					1	2	3	1	2	3
Fourier	–	20	14.786	13.470	61.9	52.0	51.1	61.0	55.0	51.7
WMVDR	60	20	14.487	14.161	60.9	51.2	49.7	59.6	51.7	49.5
WMVDR	30	20	15.111	14.155	59.0	50.5	48.9	59.3	52.1	49.9
W2MVDR	60	20	15.380	14.116	60.3	51.1	49.8	59.9	50.4	47.9

Channel Combination and Selection
Class Separability Measure

- Separate between eight classes on speech frames (silence frames not considered)
- Calculate the within-class S_w^{ch} and between-class S_b^{ch} scatter matrices for each channel ch
- Calculate class separability and chose channel with maximal separability

$$ch = \arg\max_{ch} trace\left\{(W^T S_w^{ch} W)^{-1}(W^T S_b^{ch} W)\right\}$$

Channel Selection	Word Error Rate %		
Pass	1	2	3
Signal to Noise Ratio	73.0	62.3	59.5
Class Separability Measure	67.4	57.8	55.1

Delay and Sum vs. Channel Selectionon on RT-07 eval. shows a relative improvement on the second pass of 3.6%, from 52.4% to 50.5% WER

Acoustic Model Training

•The training setup was based on last years evaluation system:

- Semi-continious quint phone system
- 16000 distributions over 4000 codebooks, max. 64 Gaussians per model

• Trained on

- Meeting Data: CMU, ICSI, NIST + Phase 2 Part 1
- Lecture Data: TED, and CHIL + RT-06S lecture meeting dev and eval

• Far-field adaptation (4 Viterbi, MAP) performed on CHIL tabel-top data.

• In addition to last years system the models have been improved by MMIE training

Language Model, Dictionary and Lexicon

• Language Model

- Similar to last year, fine tuning with new data on a new tuning set gave improvements up to 0.4% WER
- Perplexity on RT-07 dev. 123 and eval. 101

• Dictionary

- Same as last year, OOV on RT-07 dev. 0.7% and eval. 0.6%

• Lexicon

- Same as last year

WER and RT Factors

RT07 development and evaluation

condition	IHM		SDM	MDM		
pass	dev	eval	eval	dev	compare	eval
1	36.5	43.1	57.9	56.7	60.2*	56.5
2	29.5	36.3	54.9	50.5	56.8	52.4
3	28.6	36.7	54.4	49.4	54.4	52.1
RT		91	113			114

adaptation is not working as supposed to be
due to adaptation increase of deletions up to 4%

* has been adapted incrementally

Thank you for your attention!

Time for questions ...