UT-Scope: Towards LVCSR Under Lombard Effect Induced by Varying Types and Levels of Noisy Background

Hynek Boril and John H.L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
Department of Electrical Engineering
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.

ICASSP 2011

May 22-27, 2011 Prague, Czech Republic

Outline

- Introduction
- UT-Scope Database
- Speech Production Under Lombard Effect (LE)
- Modified RASTA Filter for ASR
- QCN_RASTA Normalization
- LVCSR Evaluation
- Conclusions

Introduction

What is Lombard Effect?

- Communication in noisy environments → speakers adjust their speech production in effort to maintain intelligible communication (= Lombard effect, LE)
- LE is represented by increase of vocal effort, increase of pitch, shifts of low formants, formant bandwidth reduction, spectral slope flattening, ...
- ASR acoustic models trained typically on neutral speech → ASR deterioration in LE (mismatch between acoustic models and LE speech parameters)

Objective

- Previous ASR studies mostly focused on LE in small vocabulary tasks
  → Focus on LE in large vocabulary continuous speech data
- Analysis of LE speech production in UT-Scope database
- Proposal of temporal filtering strategy derived from RASTA
- Evaluation of state-of-the-art front-end compensations in LVCSR under LE

UT-Scope Database

- UT-Scope: Speech produced under cognitive and physical stress, emotions, and LE
- Lombard portion: 58 subjects (31 native speakers of US English – 25 F, 6 M)
- Neutral (clean) and simulated noisy conditions
- Noisy conditions: background noise samples produced through open-air headphones
- Three types of noise – car (65 mph on highway, windows half open), large crowd, pink
- Noises produced to subjects at 70, 80, and 90 dB SPL (car, crowd) and 65, 75, 85 dB SPL (pink)
- Recording in ASHA certified sound booth
- 3 microphone channels – throat mic, close-talk, and far-field mic

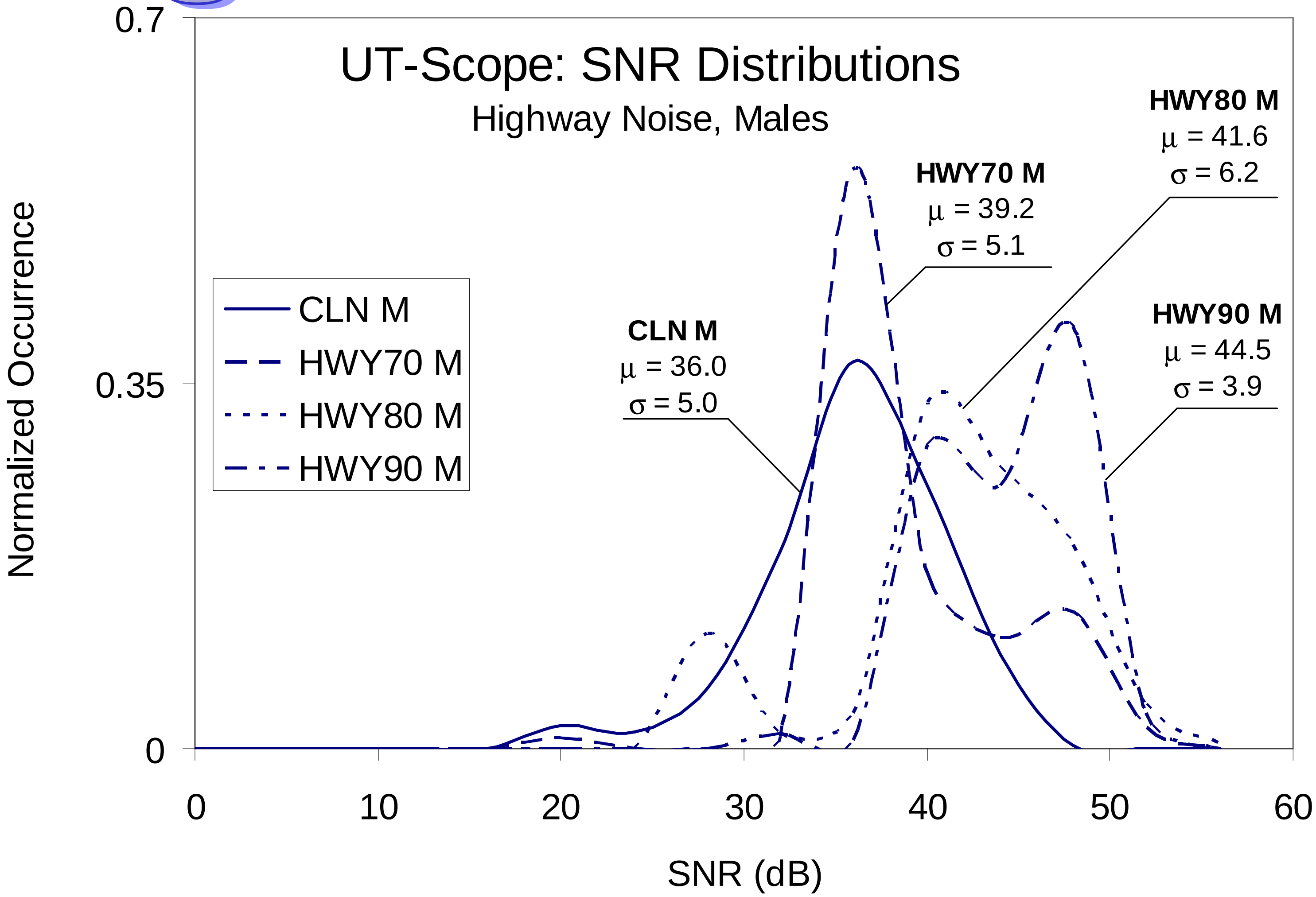
Content

- 100 phonetically balanced read sentences from TIMIT – neutral (clean) conditions
- 20 TIMIT sentences read per each of 9 noise type/level conditions
- Digit strings – 5 repetitions of 10-digit strings per each condition
- Spontaneous speech: ~1 minute per condition – describing content of a picture

Speech Production Under Lombard Effect

- Focus on TIMIT-like sentences recorded by close-talk channel
- Parameters analyzed:
  - Signal-to-noise ratio (SNR) – related to vocal intensity
  - Mean fundamental frequency (F0)
  - Vowel formant frequencies
  - Vowel durations
  - Cepstral distributions
- Extraction tools:
  - WaveSurfer (F0, formants)
  - Segmental SNR estimation tool (CTU in Prague)
  - HTK – forced alignment (vowel boundaries in formant analysis, vowel durations)
  - CTU Copy – extraction of cepstral features

Signal-to-Noise Ratio



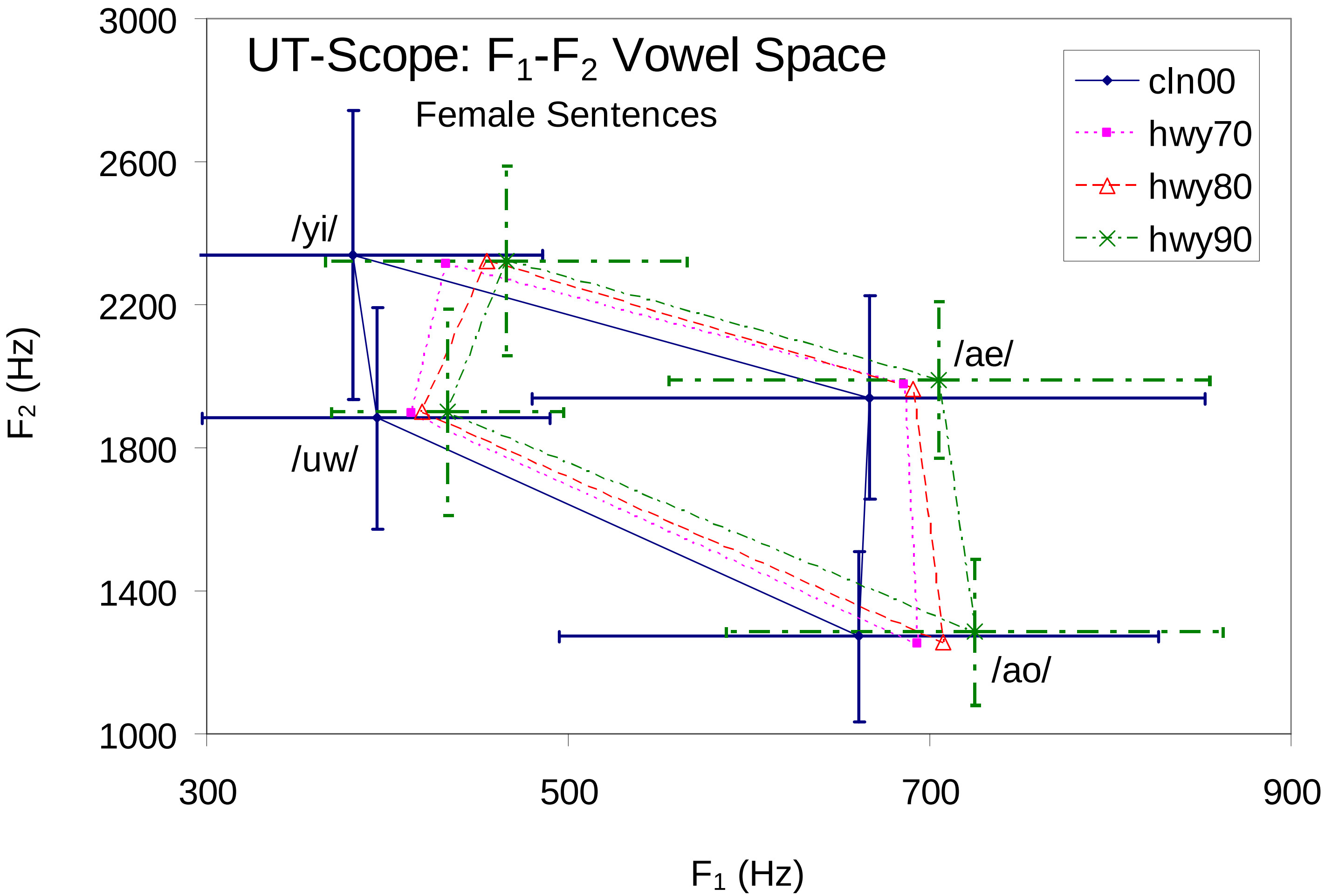
- **Lombard function** (LF) – relation between noise level and speech intensity
- Subjects increase vocal effort with the level of noise; **observed LF slopes** here **0–0.3**

Mean Fundamental Frequency (F0)

	HWY (dB)			CRD (dB)			PNK (dB)		
Gend	70	80	90	70	80	90	65	75	85
F	a=0.938, R^2=0.999 MSE=0.068			a=0.808, R^2=0.998 MSE=0.083			a=0.596, R^2=0.984 MSE=0.380		
M	a=1.195, R^2=1.000 MSE=0.039			a=1.073, R^2=1.000 MSE=0.011			a=0.786, R^2=0.962 MSE=1.634		

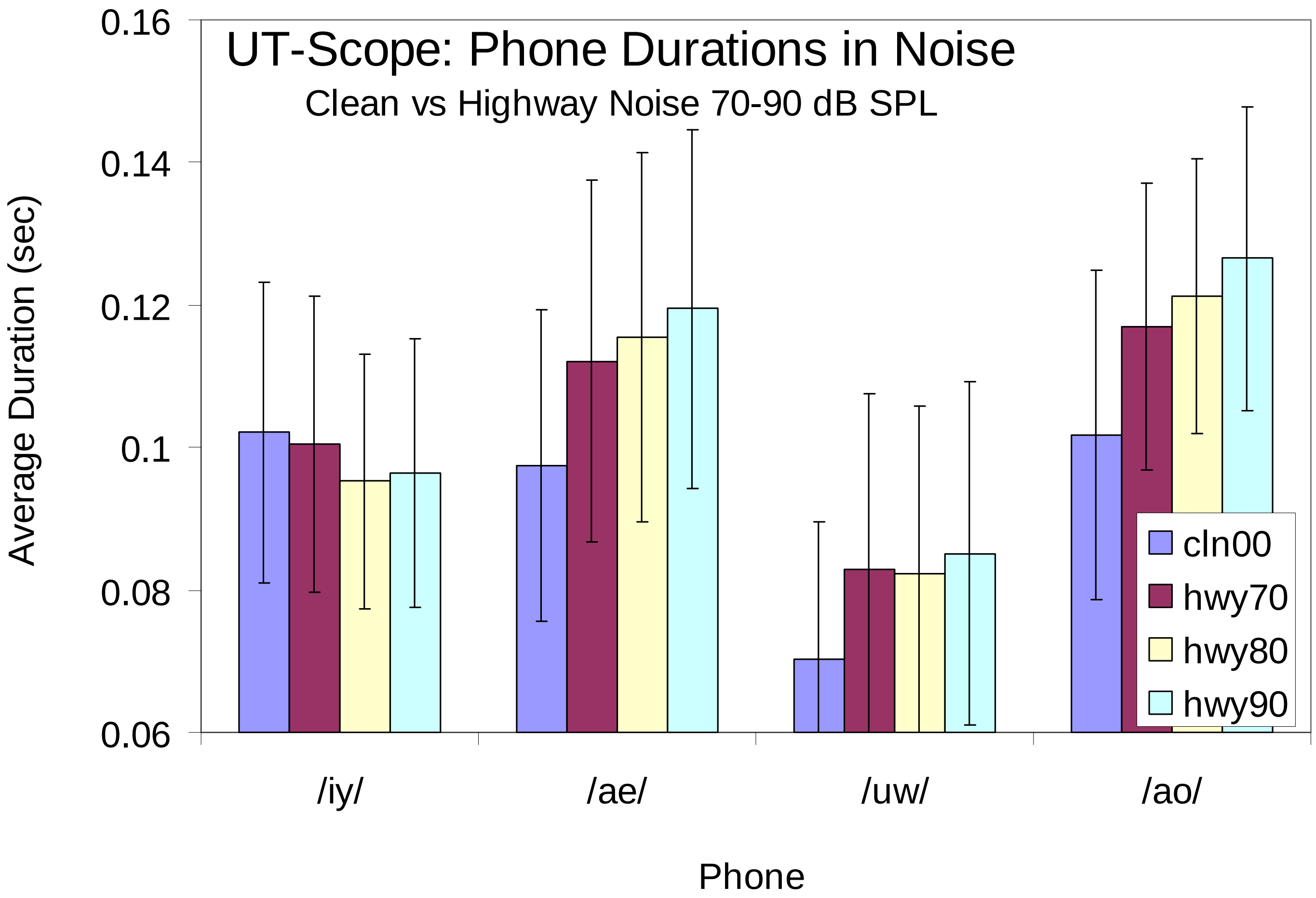
- **Correlation** analysis **between noise presentation level** (in dB) **and mean F0** across all recordings in that noise level
- a – slope of the regression line in the noise level/F0 plane; R^2 – correlation coefficient; MSE – mean square error
- Consistent F0 increase with the level of noise; steepest for car noise
- R^2, MSE – strong correlation between the level of noise and speech intensity (hwy, pnk)

Vowel Formant Frequencies



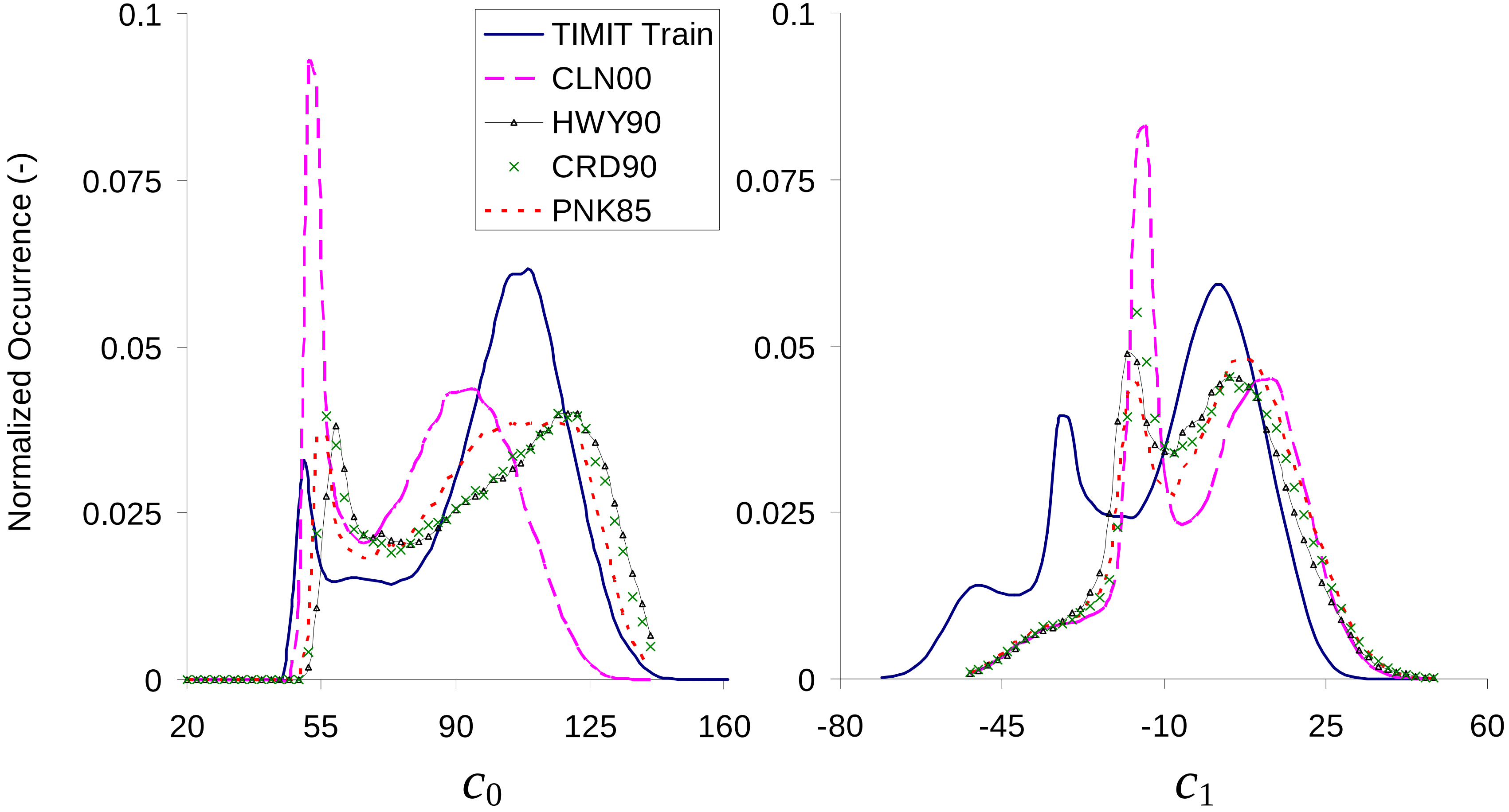
- **Vowel** segment **boundaries** estimated through **forced alignment**
- Systematic shift of vowels in F1-F2 space with increasing noise level

Vowel Durations



- Vowel segments estimated through forced alignment
- **Increasing trend** in some vowel durations, not statistically significant (95% CI's)

Cepstral Distributions



- **Speech production variations in LE** – direct impact on ASR features (here c0, c1 in MFCC) – these plots are for **clean speech signal** (high SNR)
- **Channel differences** – another source of mismatch – compare **TIMIT and CLN00**

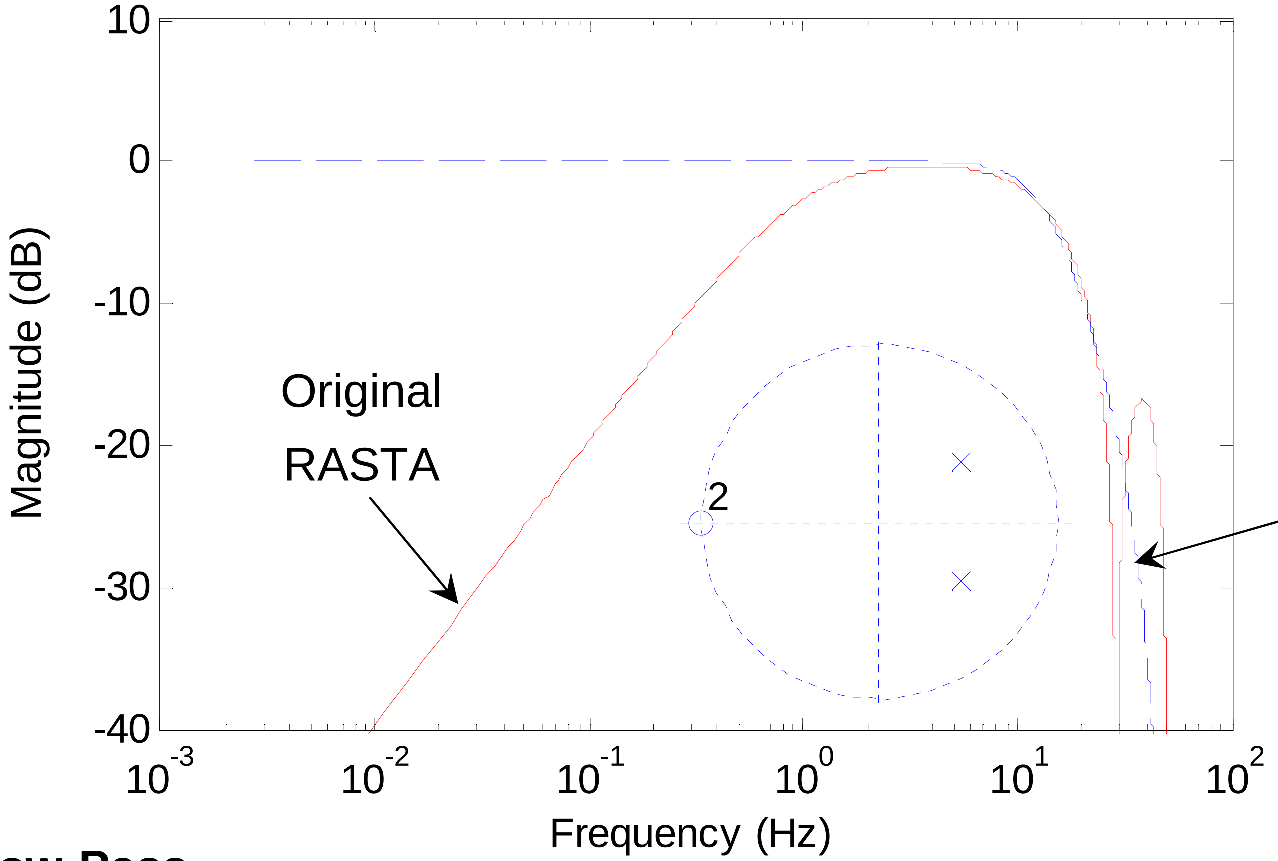
Modified RASTA Filter for ASR

- RASTA– band-pass filtering in log-spectral or cepstral domain; elimination of slow-varying components (including DC) and components varying faster than expected for speech
- RASTA is popular in ASR and speaker ID as it increases robustness to channel variations, reverberation, and noise
- Original RASTA filter - high order IIR band-pass filter – introduces transient distortions in the feature tracks

Proposed Modification

- RASTA can be approximated by a combination of cepstral mean normalization (CMN) and a low-pass filter, i.e., by **distribution normalization** & **temporal filtering**
  → decomposition of RASTA into two blocks
  → low-pass – requires lower order filter → reduced transient effects
  → allows for replacement of first block (CMN) by more powerful normalizations

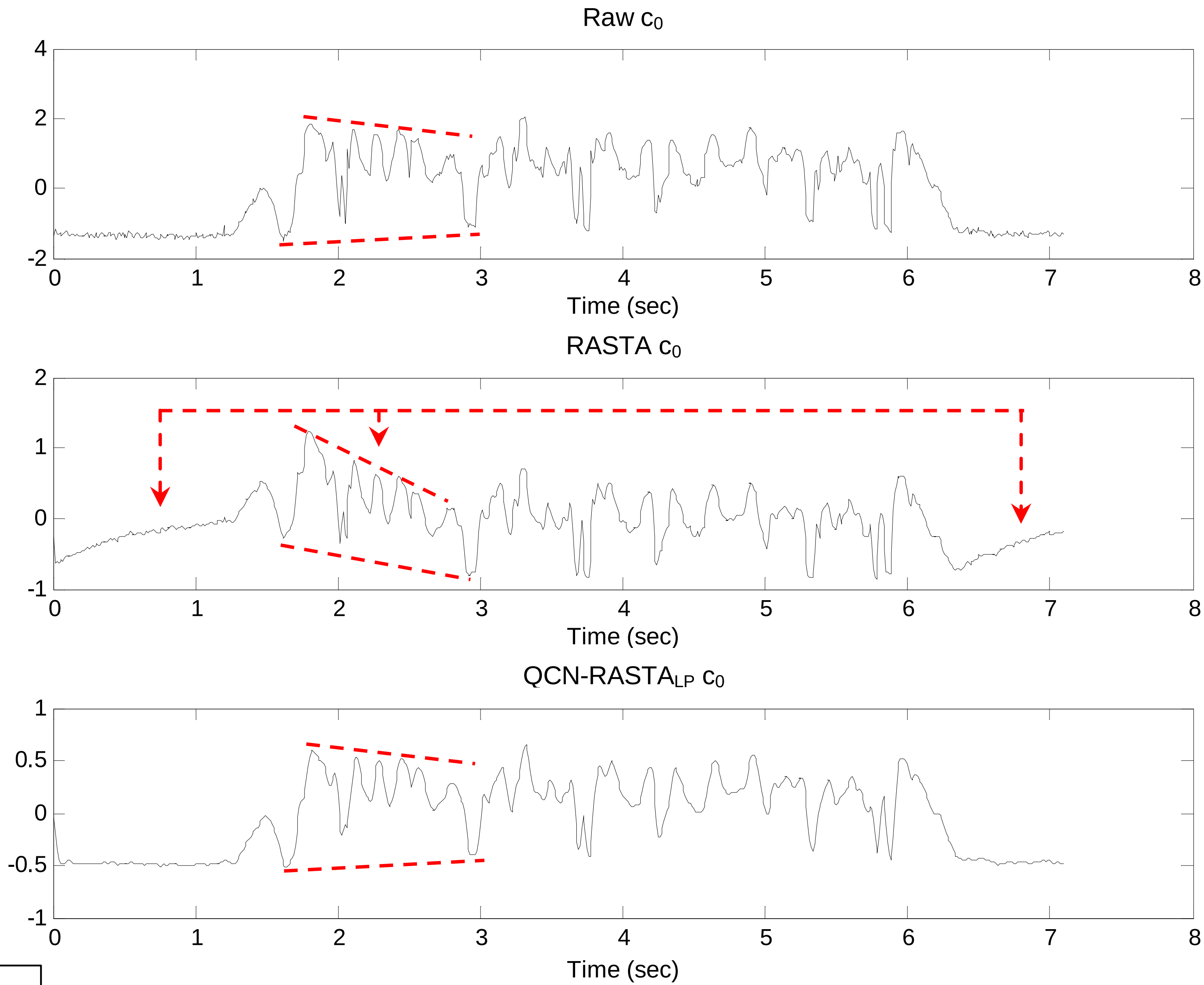
Modified RASTA Filter for ASR



Proposed Low-Pass

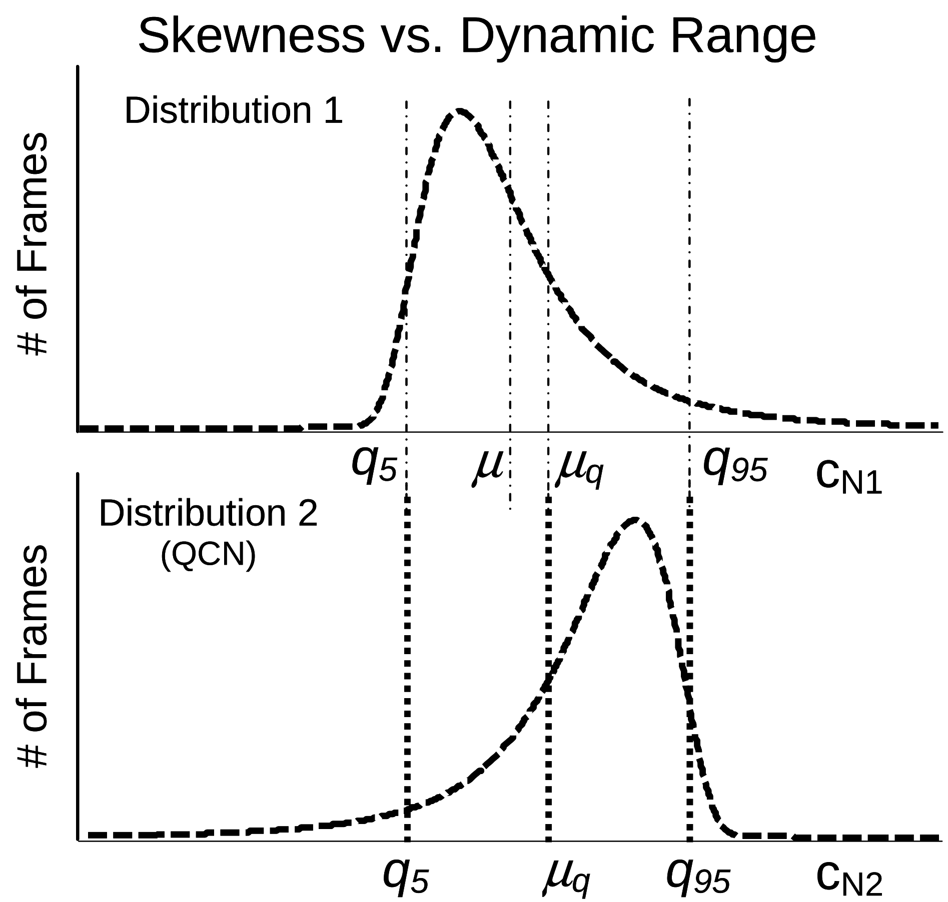
- 2nd order low-pass IIR filter (Butterworth approximation)
- Transfer function is smooth – **eliminates the residual side lobe of original RASTA**

Modified RASTA Filter for ASR



QCN_RASTA Normalization

- QCN – quantile-based cepstral dynamics normalization – introduced at ICASSP'09
- QCN aligns dynamic ranges rather than means of cepstral distributions – found to provide better normalization of distributions with **different skewness due to noise & LE**
- QCN_RASTA – QCN (replacing CMN) + proposed low-pass filter

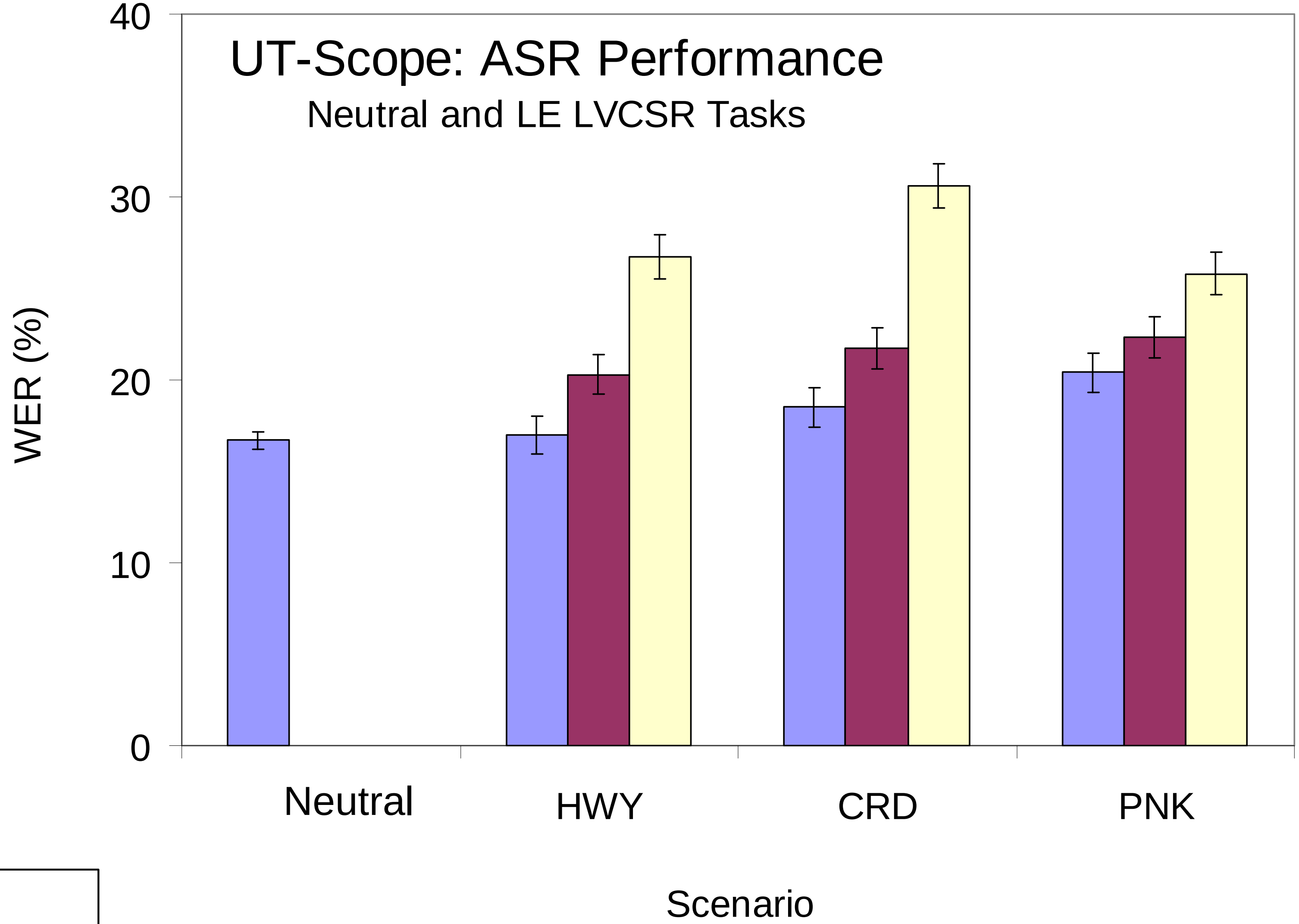


LVCSR Evaluation

- ASR system:
  - acoustic model – triphone HMM's, 32 mixtures (HTK); trained on clean TIMIT
  - language model – SRI LM Toolkit
  - TIMIT acoustic models adapted towards UT-Scope using MLLR and MAP; adapt set - 9 UT-Scope sessions (excluded from open test set)
  - clean test set – 3 male and 9 female subjects, 1 neutral and 9 simulated noisy conditions per subject
  - noisy test set – neutral speech and speech produced in 90 dB of highway noise – both mixed with NOISEX'92 Volvo noise at 5 dB and 15 dB SNR (3 M, 9 F)
- Baseline ASR performance on clean neutral test set
  - MFCC-CVN front-end: 8.3% WER
  - PLP-CVN front-end: 8.9% WER
  - All following results reported for **MFCC**-based systems with **LM off**

LVCSR Evaluation

- Impact of LE on LVCSR: clean recordings (no noise added); MFCC-CVN front-end; no LM
- WER systematically increases with the level of LE



LVCSR Evaluation

- Evaluation of selected cepstral compensation strategies:
  - Cepstral mean normalization (CMN)
  - Cepstral mean-variance normalization (CVN)
  - Cepstral gain normalization (CGN)
  - RASTA filtering in cepstral domain
  - Feature warping (Gaussianization on the utterance level)
  - Histogram equalization (TIMIT training data → reference distributions)
  - Quantile-based cepstral dynamics normalization (QCN); QCN4 – 4% and 96% quantiles bound the dynamic range; QCN9 – utilizes 9% and 91% quantiles
  - QCN_RASTA – QCN + proposed low-pass filter

LVCSR Evaluation

Clean Recordings	
Cepstral Comp.	Across Cond.
none	62.0
RASTA	60.0
warp	55.7
CMN	54.3
QCN4	54.3
HistEq	**53.9**
CVN	**53.3**
CGN	**52.8**
QCN4_RASTA	**52.6**
QCN9	**51.1**

Noisy Recordings	
Cepstral Comp.	Across Cond.
none	77.8
QCN9	69.2
CVN	68.5
QCN4_RASTA	68.4
CGN	67.0
HistEq	64.4

- Proposed QCN_RASTA improves performance of QCN; QCN-normalized features outperform other considered setups on clean neutral and LE recordings
- The ranking of front-ends changes in noisy conditions (recordings mixed with car noise); QCN_RASTA still outperforms 'raw' QCN normalization

Conclusions

- **Analyzed** impact of **LE** on **speech parameters** in UT-Scope database
- A number of speech production **parameters** found to **vary** with the type and level of noise inducing LE
- Strong **linear relationship** between **noise presentation level** (dB) and **mean pitch** (Hz) was observed for large crowd and highway noises
- A **modified** version of **RASTA** filtering scheme was proposed and shown to **reduce transient effects** of original RASTA
- **Combination of QCN** and newly designed **low-pass filter** (**QCN_RASTA**) improved QCN performance in both clean signal and noisy signal conditions (on a mixture of neutral and LE speech)
- A number of **cepstral normalizations were compared** in the task of talking style (neutral/LE) and noisy background mismatch
- CGN, histogram equalization, QCN, and newly proposed QCN_RASTA provided **significant performance gains** in talking style/noise mismatched conditions
- None of the normalizations managed to provide superior performance across all tasks

Thank you!