TOWARDS MICROPHONE-INDEPENDENT SPEECH RECOGNITION

Alejandro Acero
Richard M. Stern

Department of Electrical and Computer Engineering
and School of Computer Science
Carnegie Mellon University

BASELINE PERFORMANCE AND GOALS

Baseline performance

TRAIN TEST	CLSTK CLSTK	CLSTK CRPZM	CRPZM CLSTK	CRPZM CRPZM
BASE	85.3%	18.6%	36.9%	76.5%

Immediate challenges:

- Improve performance for cross conditions (robustness)
- Improve absolute CRPZM/CRPZM performance

Ultimate goal: **Microphone-independent** system

- Works well with a standard microphone and acceptably with the rest.
- Does not need data about the new microphone/environment
- Works in an uncontrolled environment.
- No system does this at the present time

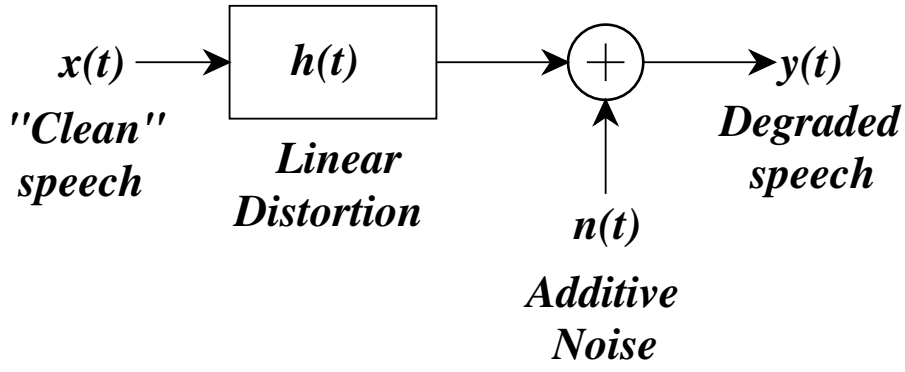
MULTI-STYLE TRAINING

TRAIN	CLSTK	CRPZM	MULTI
Test **CLSTK**	85.3%	36.9%	78.3%
Test **CRPZM**	18.6%	76.5%	69.7%

- Used in speaker independence, provides greater robustness (for "cross" conditions), but limits performance
- Better performance expected if we had a model for the degradation

A MODEL OF THE ENVIRONMENT

- Degraded speech is formed by passing "clean" (reference) speech through a filter and adding independent noise
- **Goal:** Find the parameters that undo these transformations



INDEPENDENT COMPENSATION FOR NOISE AND FILTERING

Spectral Equalization

- EQUAL. (Stockham)

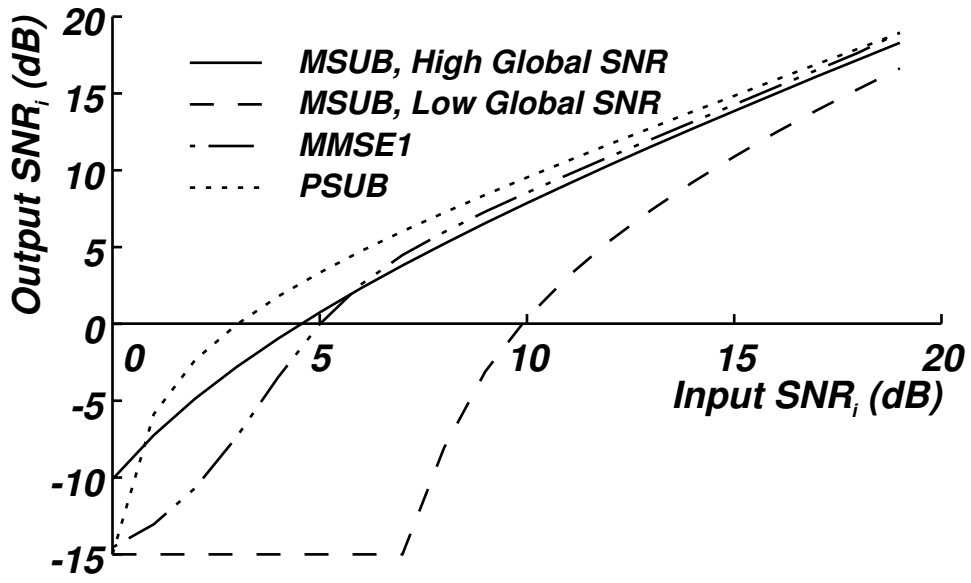
Noise suppression techniques

- **PSUB** - Boll's Power spectral subtraction
- **MSUB** - Magnitude Spectral Subtraction
- **MMSE1** - Use a transformation curve that minimizes squared error between CLSTK and PZM

[ALEX - YOU PROBABLY COULD BE EVEN A LITTLE MORE VERBOSE HERE. DO EITHER MSUB OR MMSE1 RELATE TO BEROUTI? PORTER AND BOLL? IF SO, REFERENCE THEM, TOO.]

[

INDEPENDENT COMPENSATION FOR NOISE AND FILTERING

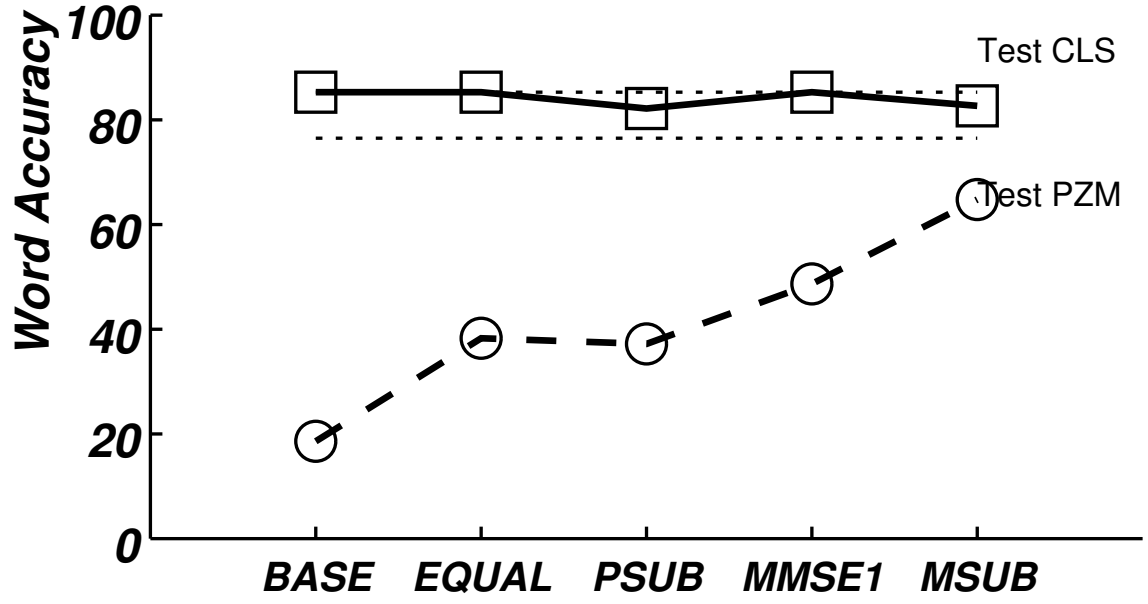


Observations

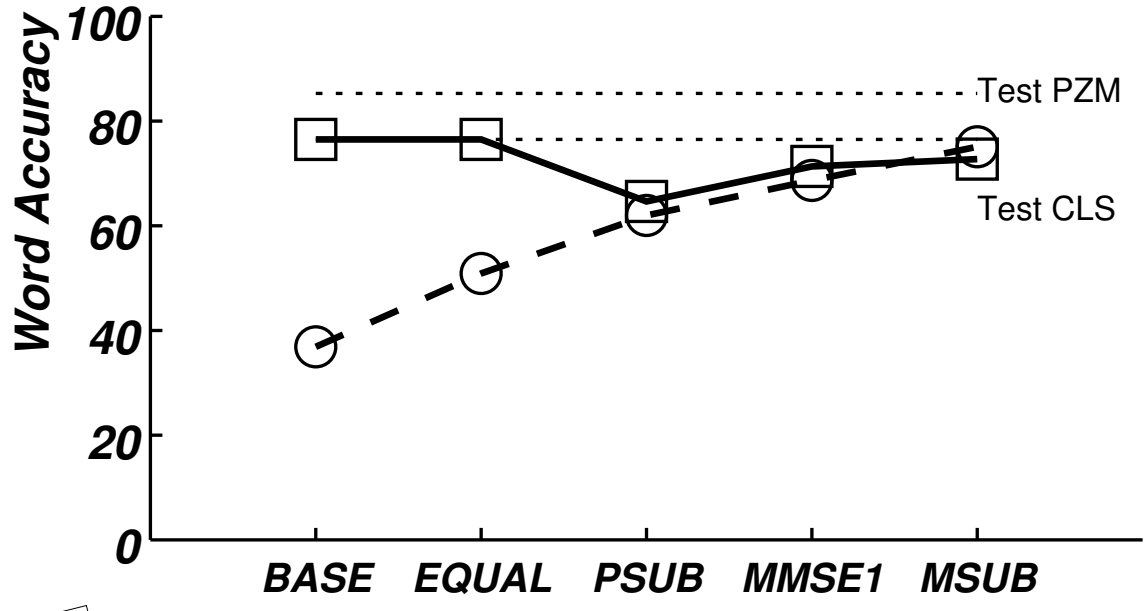
- Spectral Subtraction and Spectral Equalization interact non-linearly so a simple cascade of these algorithms does not work
- By treating different frequencies independently we obtain frames that are not speech-like

PERFORMANCE OF COMPENSATION SCHEMES

Training on Close-talking Microphone:



Training on Crown PZM Microphone:



SDCN ALGORITHM

SNR-Dependent Cepstral Normalization

w is chosen to minimize the mean-squared average difference between CLSTK and CRPZM cepstra for each SNR

Interpretation of **w**:

- Equalization at high SNR
- Noise subtraction at low SNR

TRAIN TEST	CLSTK CLSTK	CLSTK PZM	CRPZM CLSTK	CRPZM CRPZM
BASE	85.3%	18.6%	36.9%	76.5%
MMSEN	85.3%	66.4%	75.5%	72.3%
SDCN	85.3%	67.2%	76.4%	75.5%

SDCN ALGORITHM

Advantages

- *Joint* Compensation for noise and spectral tilt
- Very easy to implement, only c_0 and c_1 need to be normalized.

Disadvantages

- For every new microphone/environment, a new stereo database is needed to estimate the corresponding **w** vectors, hence
- Not microphone independent

CDCN ALGORITHM

Codeword-dependent Cepstral Normalization

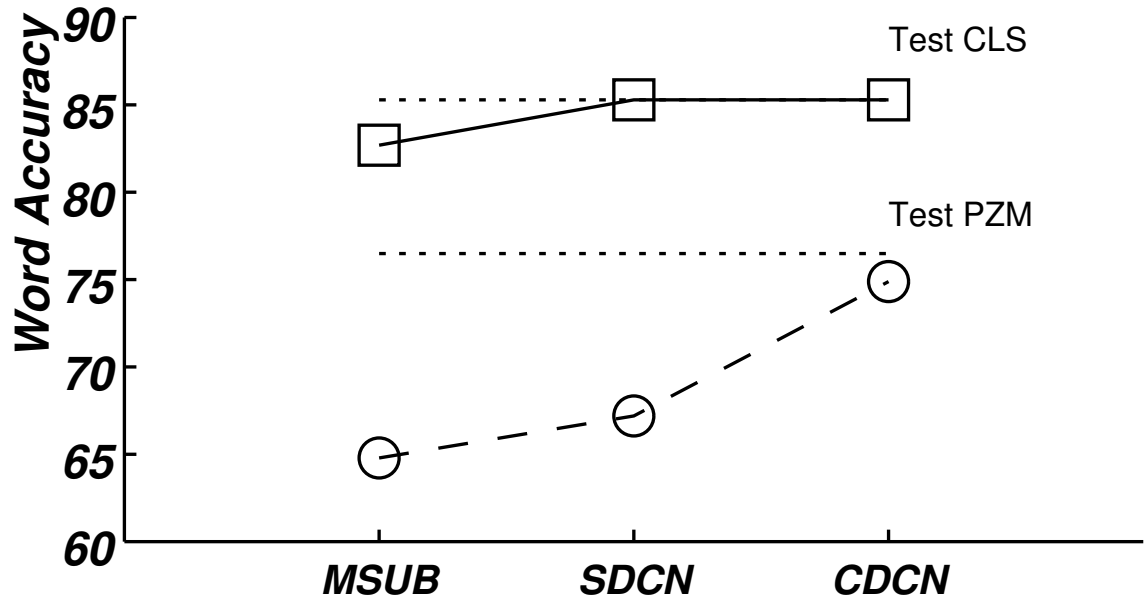
Estimation process:

1. ML estimate of **q** and **n**. Find the parameters of the transformation that maximize the probability or alternatively minimize the overall VQ distortion. Use of the EM algorithm for convergence

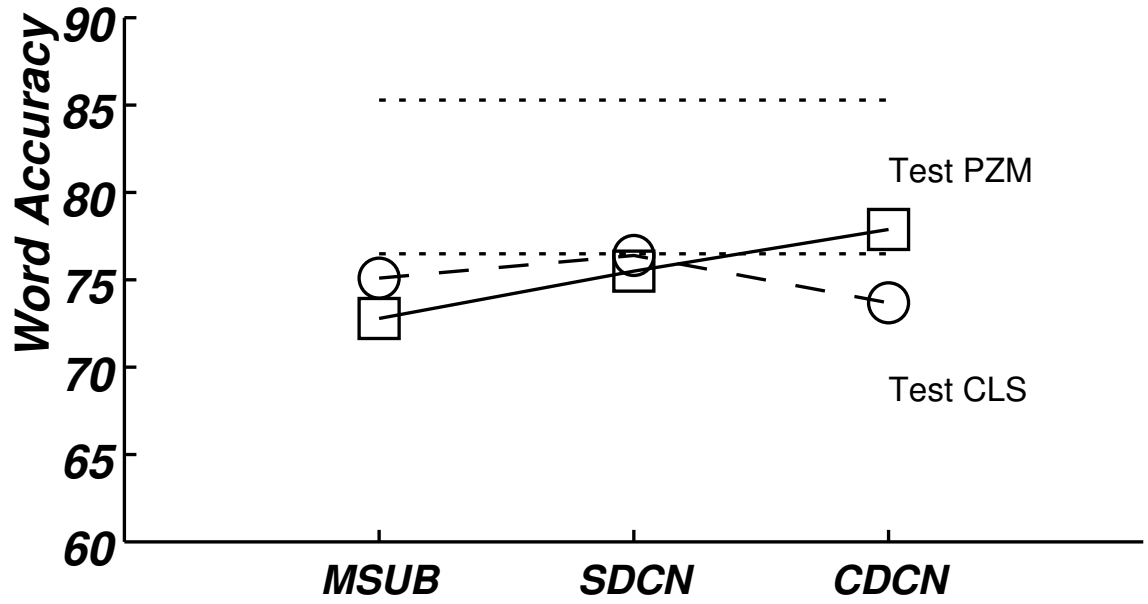
2. MMSE estimate of every cepstrum vector given **q** and **n**

PERFORMANCE OF COMPENSATION SCHEMES

Training on Close-talking Microphone:



Training on Crown PZM Microphone:



BASELINE SPECTRA

SDCN SPECTRA

CDCN SPECTRA

SPECTRAL TILT COMPARISON

CROSS MICROPHONE RECOGNITION

Performance using different microphones. In each case SPHINX had been trained with the CLSTK microphone

	CLSTK	CRPCC160
BASE	82.4%	70.2%
CDCN	81.0%	78.5%

	CLSTK	CRPZM6sf
BASE	84.8%	41.8%
CDCN	83.3%	73.9%

	CLSTK	SENN518
BASE	87.2%	84.5%
CDCN	82.2%	83.3%

	CLSTK	SENNME80
BASE	83.7%	71.4%
CDCN	81.5%	80.7%

	HMEFM	CRPCC160
BASE	55.9%	56.3%
CDCN	81.7%	72.2%

SUMMARY

- Desk-top microphones like the Crown PZM6fs increase the recognition error rate by allowing weak phonetic events to become confused with silences
- **Microphone-independent** systems can be built by estimating the parameters of the transformation: noise and spectral tilt
- A framework for speech normalization in the *cepstral domain* has been introduced