SIPL SIGNAL & IMAGE PROCESSING LAB

Statistical Methods for Speech Processing In Low Resource Environments

Hadas Benisty

PhD research under the supervision of

Prof. David Malah and Prof. Koby Crammer

22/10/2015

Overview

Voice Conversion

Global Variance Enhancement

Grid-Based Conversion

Keyword Spotting

Voice Conversion

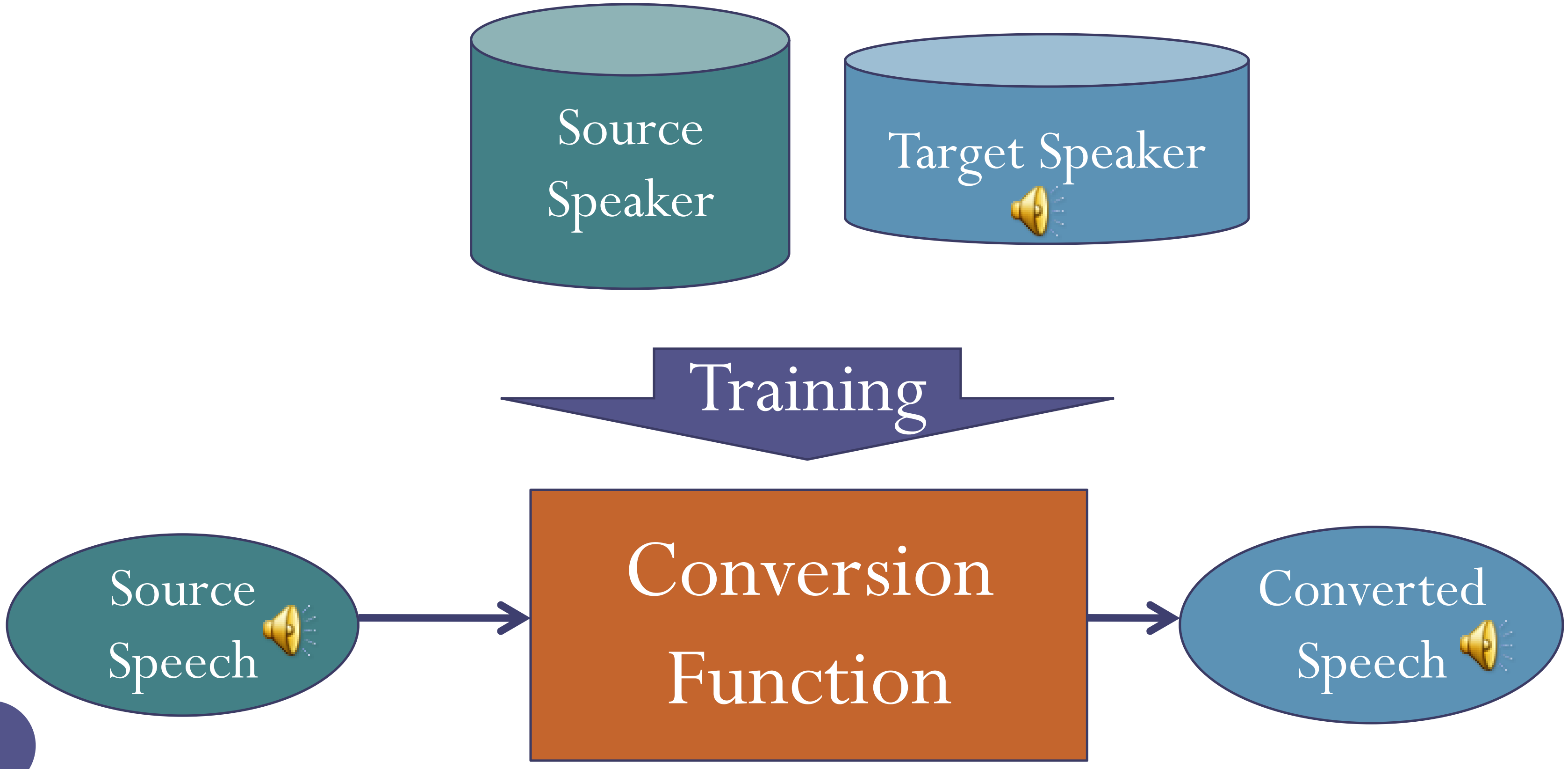
Global Variance Enhancement

Grid-Based Conversion

Keyword Spotting

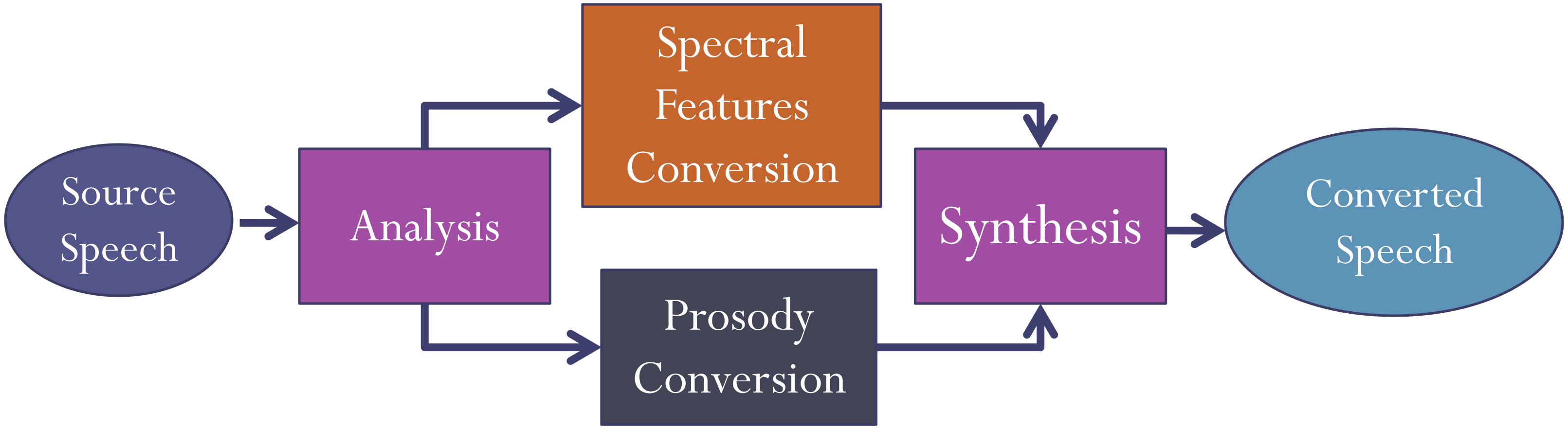
General Conversion Setup

- **The goal**: modify a source speaker's speech to sound as if spoken by a target speaker



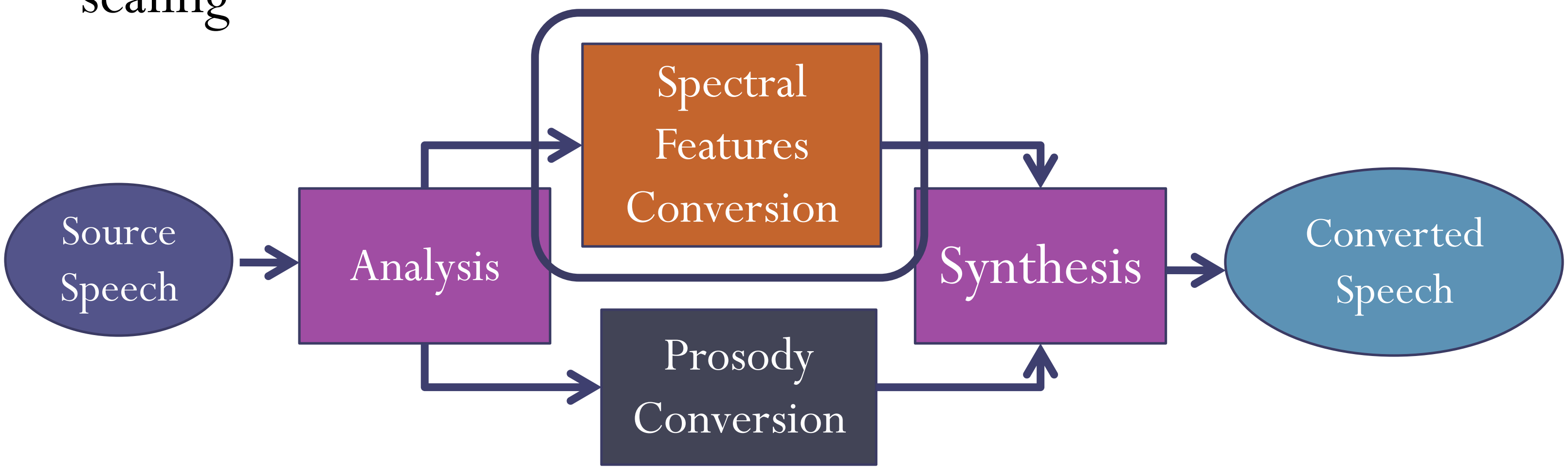
Speech Characteristics

- The identity of a speaker is associated with:
  - Prosody attributes - pitch, duration and energy
  - Spectral envelope
- Pitch - usually modified using a simple statistical mean and variance scaling



Speech Characteristics

- The identity of a speaker is associated with:
  - Prosody attributes - pitch, duration and energy
  - Spectral envelope
- Pitch - usually modified using a simple statistical mean and variance scaling

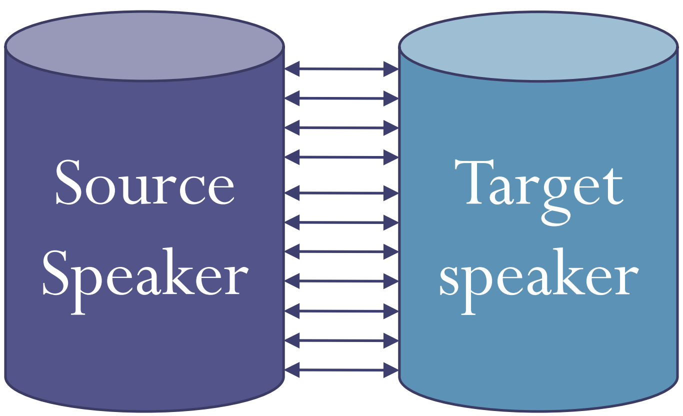


- Most VC methods deal with **spectral envelope conversion**

Training Data

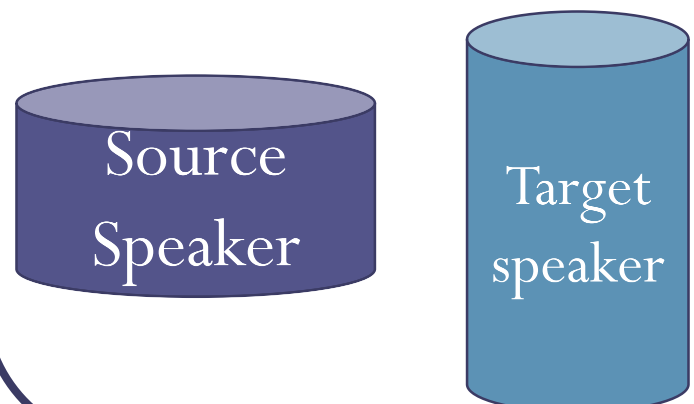
Parallel

- The source and target training sets include recordings of the two speakers say the same text



Non-Parallel

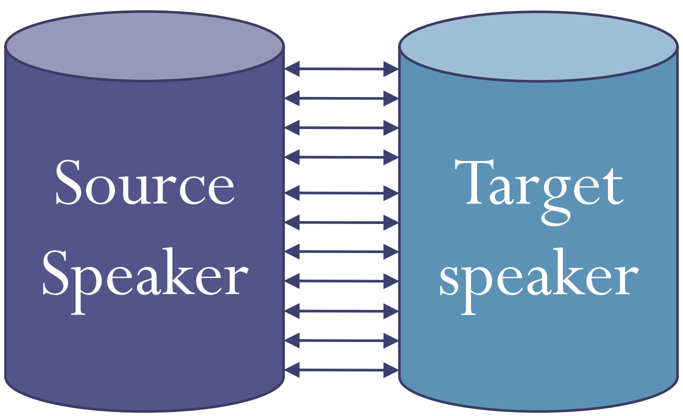
- No correspondence regarding the textual content of the training data sets is assumed



Training Data

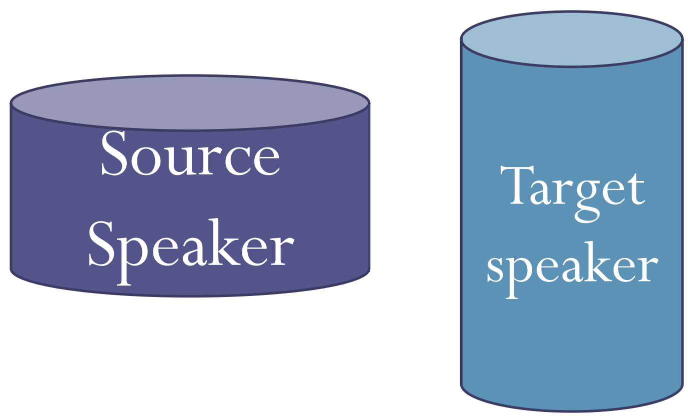
Parallel

- The source and target training sets include recordings of the two speakers say the same text



Non-Parallel

- No correspondence regarding the textual content of the training data sets is assumed



Classical GMM Conversion

Stylianou et al., 1998

- Given a parallel and aligned source and target training vectors $\{\mathbf{x}^k, \mathbf{y}^k\}_1^N \in \Re^P$ (represented by Mel Frequency Cepstrum Coefficients - MFCCs)
- A GMM is trained using the source vectors:

$$p(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m N(\mathbf{x}; \boldsymbol{\mu}_{m,} \boldsymbol{\Sigma}_m)$$

- The conversion function - a weighted sum of linear Bayesian estimators of the target spectra:

$$\boxed{\mathcal{F}(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m \left(\boldsymbol{\Gamma}_m \boldsymbol{\Sigma}_m^{-1} (\mathbf{x} - \boldsymbol{\mu}_{m,}) + \boldsymbol{\nu}_m \right)}$$

Classical GMM Conversion – Cont'd

- The conversion parameters $\mathbf{v}_m, \mathbf{\Gamma}_m$ - evaluated using Least Squares
- Minimizing the mean spectral distance between the **converted** and **target** spectra:

$$\boxed{\min_{\substack{\mathbf{v}_m, \mathbf{\Gamma}_m \\ m=1,\ldots,M}} \left\{ \sum_{k=1}^{N} \mathrm{MCD}^2\left(\mathcal{F}\left(\mathbf{x}^k\right), \mathbf{y}_k\right) \right\}}$$

  - where:

MCD – Mel Cepstrum Distortion:

$$\mathrm{MCD}\left(\mathcal{F}\left(\mathbf{x}^k\right), \mathbf{y}_k\right) = \frac{10\sqrt{2}}{\ln 10} \left\| \mathcal{F}\left(\mathbf{x}^k\right) - \mathbf{y}^k \right\|_2$$

Limitations of GMM-Based Conversion Methods

- **Model Selection**
  - **A high order model**
    - Over fitting → poor prediction ability on new data
  - **A low order model**
    - Over-smoothed spectral envelopes → muffled synthesized speech
- **Frame-By-Frame Conversion**

Limitations of GMM-Based Conversion Methods

- **Model Selection**
  - **A high order model**
    - Over fitting → poor prediction ability on new data
  - **A low order model**
    - Over-smoothed spectral envelopes → muffled synthesized speech
- **Frame-By-Frame Conversion**

Low Quality of Synthesized Speech

Speaker	Signal
Source	
Target	
GMM	

Limitations of GMM-Based Conversion Methods – Cont'd

- **Training Data Size**
  - Several dozen sentences
- **Iterative Training**
  - Expectation Maximization
- **Training Set**
  - Parallel sentences
  - Aligned data set (using Dynamic Time Warping (DTW))

Limitations of GMM-Based Conversion Methods – Cont'd

- **Training Data Size**
  - Several dozen sentences
- **Iterative Training**
  - Expectation Maximization
- **Training Set**
  - Parallel sentences
  - Aligned data set (using Dynamic Time Warping (DTW))

Problematic for low resource applications

Proposed Solutions

- **<u>Global Variance (GV) Enhancement</u>**
  - **Constraint GMM – INTERSPEECH 2011**
    - GMM-based conversion with a GV Constraint
  - **Modular Global Variance (GV) Enhancement - EUSIPCO 2012**
    - A modular GV enhancement method applied as a **post-processing** block

Proposed Solutions

- <u>**Global Variance (GV) Enhancement**</u>
  - **Constraint GMM – INTERSPEECH 2011**
    - GMM-based conversion with a GV Constraint
  - **Modular Global Variance (GV) Enhancement - EUSIPCO 2012**
    - A modular GV enhancement method applied as a **post-processing** block
- <u>**Sequential Estimation of Spectral Envelop**</u>
  - **Grid Based (GB) Conversion – Eilat 2014**
    - Temporal continuity
    - Unaligned source and target training sets
  - **GB Conversion For Low Resource Applications - Submitted**
    - Testing without phonetic segmentation

Voice Conversion

Keyword Spotting

Global Variance Enhancement

Grid-Based Conversion

GV Enhancement Approaches

- ML estimation of spectral trajectory of the converted spectra using
  - GMM - [Toda et al., 2007; Hwang et al., 2013]
  - HMM – [Zen et al., 2011]
- Limitations
  - High computational complexity
  - Cannot be applied in existing conversion systems
- Our Proposed Solutions
  - Constrained GMM (CGMM) – seamlessly applied in classical GMM-based systems
  - Modular GV Enhancement – a post processing block

CGMM (Interspeech 2011)

- Similarly to Stylianou et al.:

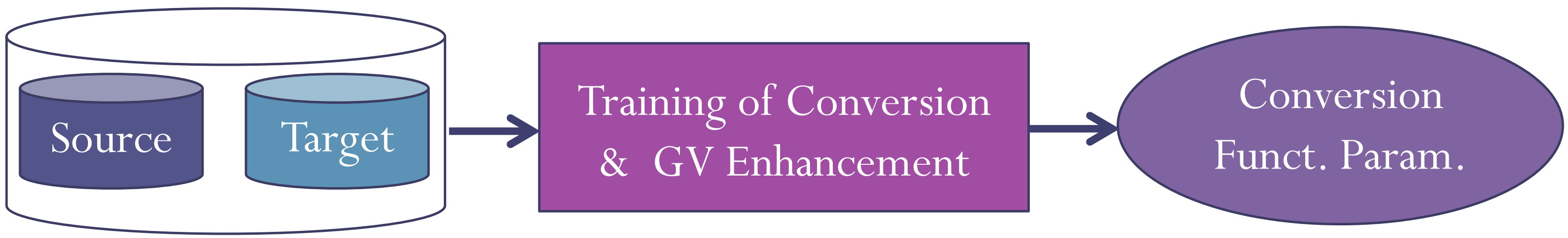
$$p(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m N\left(\mathbf{x}; \boldsymbol{\mu}_{m,} \boldsymbol{\Sigma}_m\right) \qquad \mathcal{F}(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m \left(\boldsymbol{\Gamma}_m \boldsymbol{\Sigma}_m^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_{m,}\right) + \mathbf{v}_m\right)$$

- Estimation of a linear conversion:
  - The spectral distance is minimized
  - The GV of the converted features is constrained to match the GV of the target features:

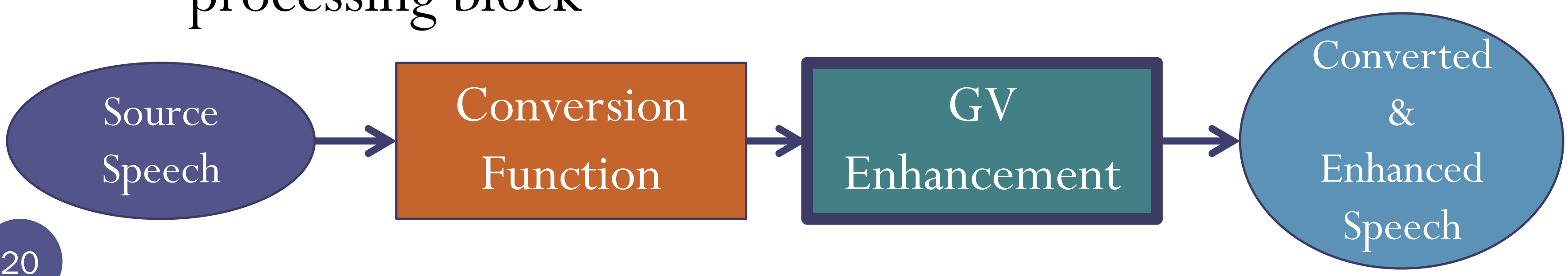
$$\boxed{\begin{array}{l} \min\limits_{\substack{\mathbf{v}_m, \boldsymbol{\Gamma}_m \\ m=1,\dots,M}} \left\{ \sum_{k=1}^{N} \mathrm{MCD}^2\left(\mathcal{F}\left(\mathbf{x}^k\right), \mathbf{y}^k\right) \right\} \\ \text{s.t.} \quad \mathrm{Var}\left\{\mathcal{F}\left(\mathbf{x}(p)\right)\right\} = \mathrm{Var}\left\{\mathbf{y}(p)\right\} \end{array}} \qquad \begin{array}{l} \mathbf{x}^k, \mathbf{y}^k \in \Re^P \\ p = 1, \dots, P \end{array}$$

Modular GV Enhancement
(EUSIPCO 2012)

- Previously proposed enhancement methods are integrated into the training process of the conversion



- Modular GV enhancement - designed independently of any specific conversion scheme and applied as a post-processing block



Modular GV Enhancement - Cont'd
(EUSIPCO 2012)

- Given:
  - $\mathbf{Y}$ - target training set
  - $\tilde{\mathbf{Y}}_{1:T}$ - a converted sequence
- The enhanced sequence $\tilde{\mathbf{Z}}_{1:T}$ is obtained by maximizing the global variance, under a spectral distance constraint:

$$\boxed{\begin{aligned}\tilde{\mathbf{Z}}_{1:T} = \quad & \arg\max_{\mathbf{Z}_{1:T}} \mathrm{NGV}\left\{\mathbf{Z}_{1:T}\right\} \\ \text{s.t} \quad & \sum_{t=1}^{T} \mathrm{MCD}\left(\mathbf{z}_t, \tilde{\mathbf{y}}_t\right) \le \theta_{MCD}\end{aligned}}$$

A threshold specified by the user

$$\mathrm{NGV}\left\{\mathbf{Z}_{1:T}\right\} \triangleq \frac{1}{P}\sum_{p=1}^{P}\frac{\mathrm{Var}\left\{\mathbf{Z}_{1:T}(p)\right\}}{\mathrm{Var}\left\{\mathbf{Y}(p)\right\}}$$

  - We numerically solve the optimization problem using Lagrange multipliers

Experiments Results

Objective Measures

- Normalized Distortion (ND) – used for comparing conversions of several source-target sets
  - Desired value: 0

$$\mathrm{ND} = \frac{\sum_{k=1}^{N} \mathrm{MCD}^2\left(\mathcal{F}\left(\mathbf{x}^k\right), \mathbf{y}^k\right)}{\sum_{k=1}^{N} \mathrm{MCD}^2\left(\mathbf{x}^k, \mathbf{y}^k\right)}$$

- Normalized Global Variance (NGV) – GV of the converted spectra, normalized with the empirical GV of the target spectra, averaged over all P elements
  - Desired value: ~1

$$\mathrm{NGV}\left\{\tilde{\mathbf{Y}}_{1:T}\right\} \triangleq \frac{1}{P}\sum_{p=1}^{P}\frac{\mathrm{Var}\left\{\tilde{\mathbf{Y}}_{1:T}\left(p\right)\right\}}{\mathrm{Var}\left\{\mathbf{Y}\left(p\right)\right\}}$$

Objective Measures

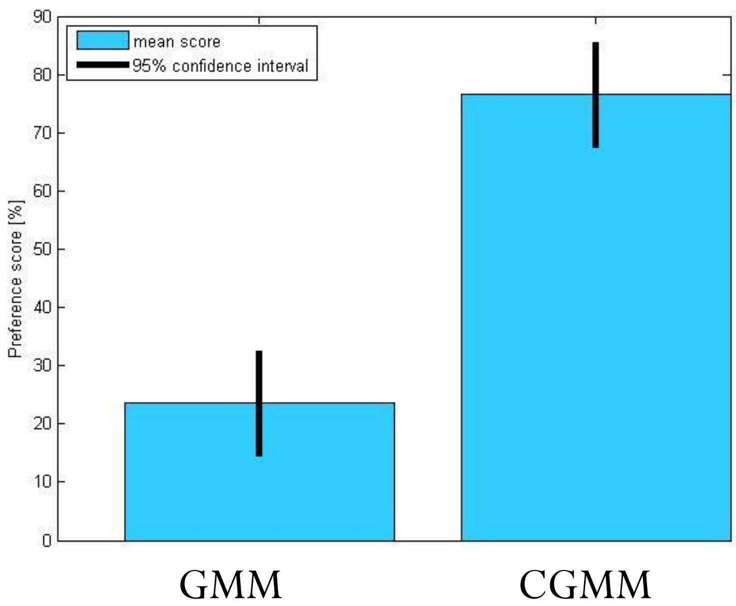
- Training set – 50 parallel and aligned sentences (male to male)
- Testing set – 50 sentences

Conversion Method	ND	NGV
GMM	0.72	0.04
GMM + Modular Enhancement ϑ=1dB	0.75	0.12
GMM + Modular Enhancement ϑ=2dB	0.78	0.15
GMM + Modular Enhancement ϑ=4dB	0.85	0.21
CGMM*	**0.85**	**0.44**

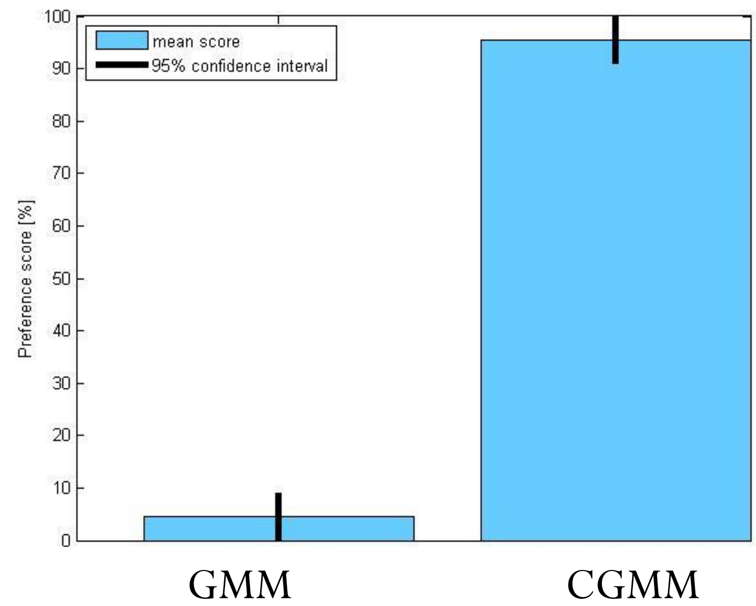
*CGMM was trained so that only the variance of the first 12 MFCCs were constraint to match the target speaker's variance

CGMM Vs. GMM

Quality Preference Test



Individuality Preference Test

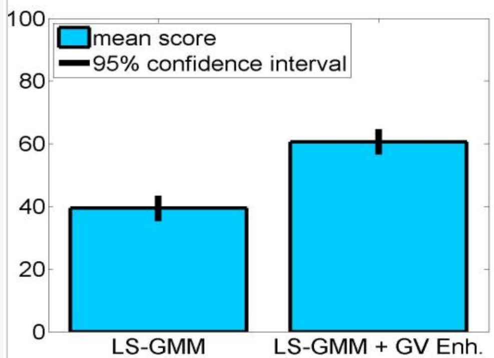


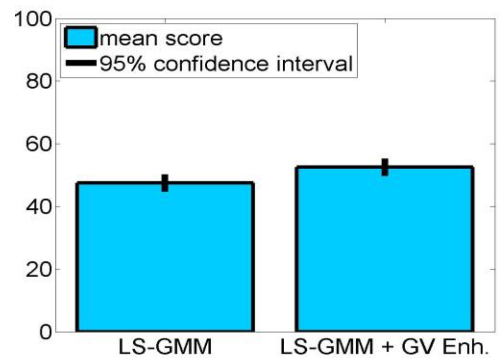
Source	Target	GMM	CGMM

Enhancement Module Vs. GMM and CGMM

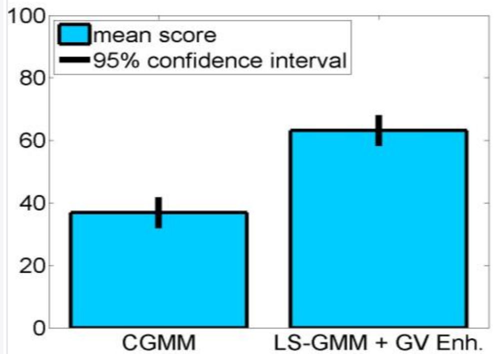
Quality | **Individuality**

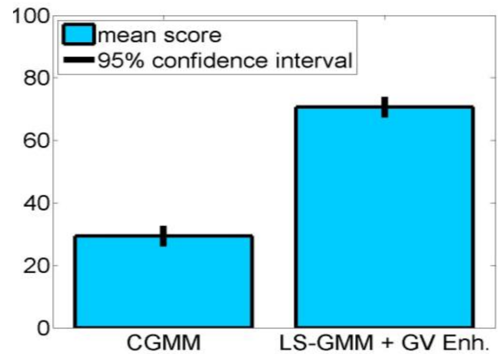
GMM vs. Enhanced-GMM





CGMM vs. Enhanced-GMM





Source	Target	GMM	CGMM	En-GMM

Voice Conversion

Keyword Spotting

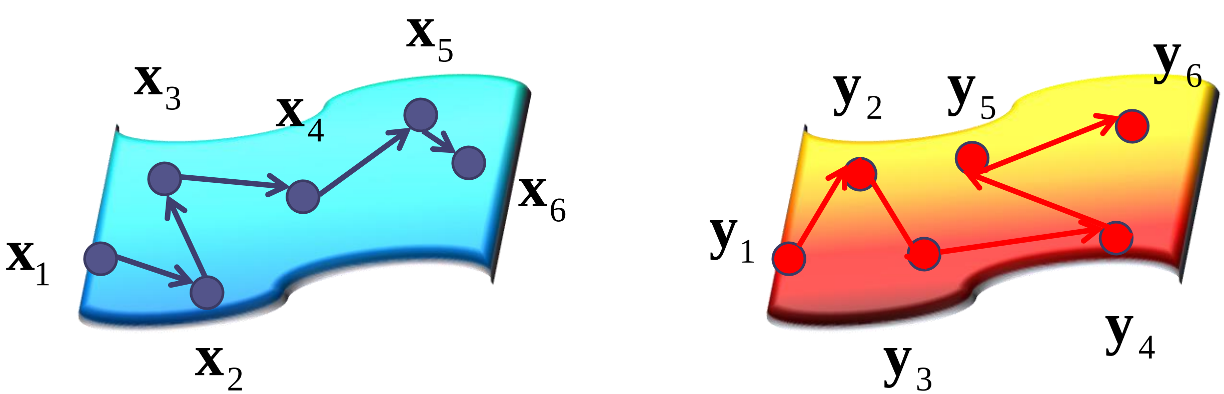
Global Variance Enhancement

Grid-Based Conversion

Grid-Based (GB) Conversion For Low Resource Applications

Main Idea

- Conversion -expressed as a **sequential estimation** problem
- The **target spectrum is tracked** based on the **observed source spectrum**

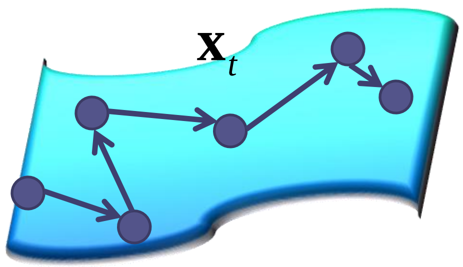


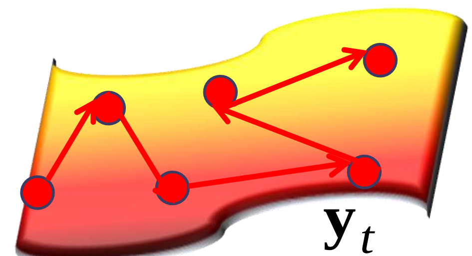
Grid-Based (GB) Conversion For Low Resource Applications – Cont'd

Advantages

- Simple non-iterative training
- Data alignment is not required (still parallel)
- Does not require phonetic segmentation at test time (unlike our initial work - IEEE-Eilat 2014)
- Trained successfully using very few sentences (5-10)

GB Conversion
Bayesian Tracking





$$\mathbf{x}_t = h_t(\mathbf{y}_t, \mathbf{v}_t) \qquad \mathbf{y}_t = f_t(\mathbf{y}_{t-1}, \mathbf{u}_t)$$

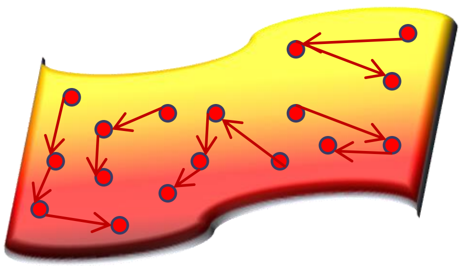
- The Bayesian optimal estimation for the target spectrum is:
$$\hat{\mathbf{y}}_t = E\left[\mathbf{y}_t \middle| \mathbf{x}_{1:t}\right] = \int p\left(\mathbf{y}_t \middle| \mathbf{x}_{1:t}\right)\mathbf{y}_t d\mathbf{y}_t$$
- In practice - analytical derivation requires modeling of
$$p\left(\mathbf{y}_t \middle| \mathbf{x}_{1:t}\right)$$
- Instead - we use a **Grid-Based approximation**

GB Conversion

Discrete Approximation

- We evaluate the posterior probability as a discrete sum:

$$p\left(\mathbf{y}_t \middle| \mathbf{x}_{1:t}\right) = \sum_{k=1}^{N_y} w_{t|t}^k \delta\left(\mathbf{y}_t = \mathbf{y}^k\right)$$



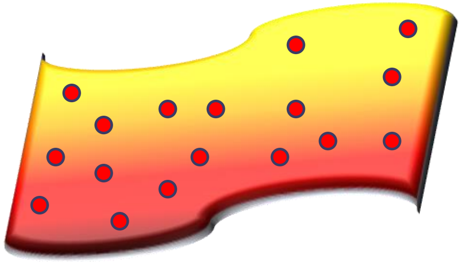
$$\left\{\mathbf{y}^k\right\}_{k=1}^{N_y}$$

GB Conversion

Discrete Approximation

- We evaluate the posterior probability as a discrete sum:

$$p\left(\mathbf{y}_t \middle| \mathbf{x}_{1:t}\right) = \sum_{k=1}^{N_y} w_{t|t}^{k} \delta\left(\mathbf{y}_t = \mathbf{y}^k\right)$$

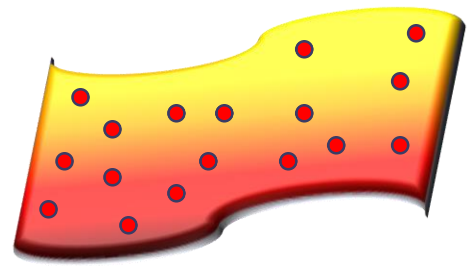


$$\left\{\mathbf{y}^k\right\}_{k=1}^{N_y}$$

GB Conversion
Discrete Approximation

- We evaluate the posterior probability as a discrete sum:

$$p\left(\mathbf{y}_t \middle| \mathbf{x}_{1:t}\right) = \sum_{k=1}^{N_y} w_{t|t}^k \delta\left(\mathbf{y}_t = \mathbf{y}^k\right)$$



$$\left\{\mathbf{y}^k\right\}_{k=1}^{N_y}$$

- The optimal Bayesian estimation - a discrete sum of the **target training vectors**:

$$\boxed{\hat{\mathbf{y}}_t = \sum_{k=1}^{N_y} w_{t|t}^k \mathbf{y}^k}$$

Where:

- The posterior weights are: $w_{t|t}^k \approx p\left(\mathbf{y}_t = \mathbf{y}^k \middle| \mathbf{x}_{1:t}\right)$
- These weights are evaluated using a parallel unaligned training sets:

$$\left\{\mathbf{x}^k\right\}_1^{N_x}, \left\{\mathbf{y}^k\right\}_1^{N_y} \in \Re^P$$

GB Conversion

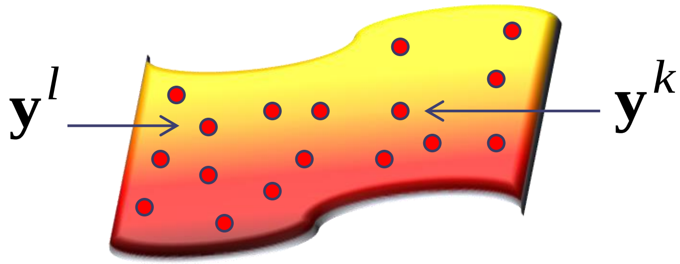
Sequential Estimation Of The Posterior Weights

- The posterior weights are sequentially evaluated using two stages:
  1. Prediction: $$w_{t|t-1}^{k} = \sum_{l=1}^{N_y} w_{t-1|t-1}^{l} \, p\left(\mathbf{y}_t = \mathbf{y}^k \middle| \mathbf{y}_{t-1} = \mathbf{y}^l\right)$$
  2. Update: $$w_{t|t}^{k} = \frac{w_{t|t-1}^{k} \, p\left(\mathbf{x}_t \middle| \mathbf{y}_t = \mathbf{y}^k\right)}{\sum_{l=1}^{N_y} w_{t|t-1}^{l} \, p\left(\mathbf{x}_t \middle| \mathbf{y}_t = \mathbf{y}^l\right)}$$

- $p\left(\mathbf{y}_t = \mathbf{y}^k \middle| \mathbf{y}_{t-1} = \mathbf{y}^l\right)$ - evidence probability
- $p\left(\mathbf{x}_t \middle| \mathbf{y}_t = \mathbf{y}^k\right)$ - likelihood probability

GB Conversion
Evidence Modeling

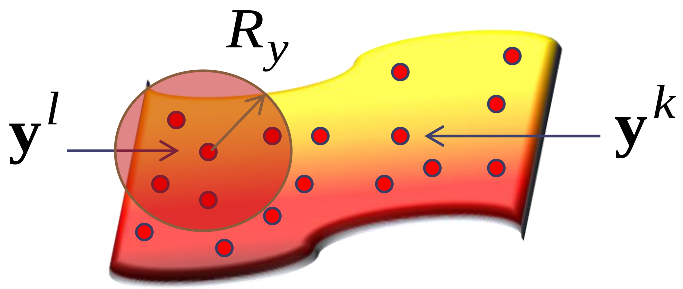
- A transition probability $\mathbf{y}^l \rightarrow \mathbf{y}^k$ at time t $p\left(\mathbf{y}_t = \mathbf{y}^k \middle| \mathbf{y}_{t-1} = \mathbf{y}^l\right)$



GB Conversion

Evidence Modeling

- A transition probability $\mathbf{y}^l \rightarrow \mathbf{y}^k$ at time t $p\left(\mathbf{y}_t = \mathbf{y}^k \middle| \mathbf{y}_{t-1} = \mathbf{y}^l\right)$

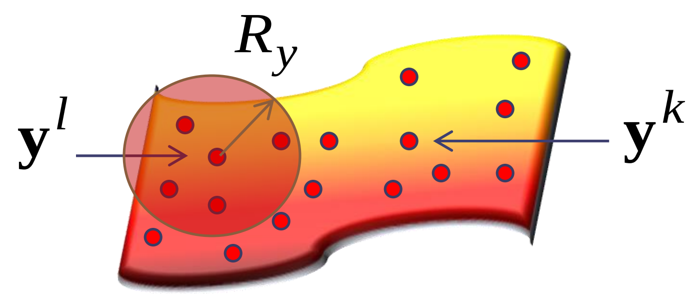


  - R_y – a parameter

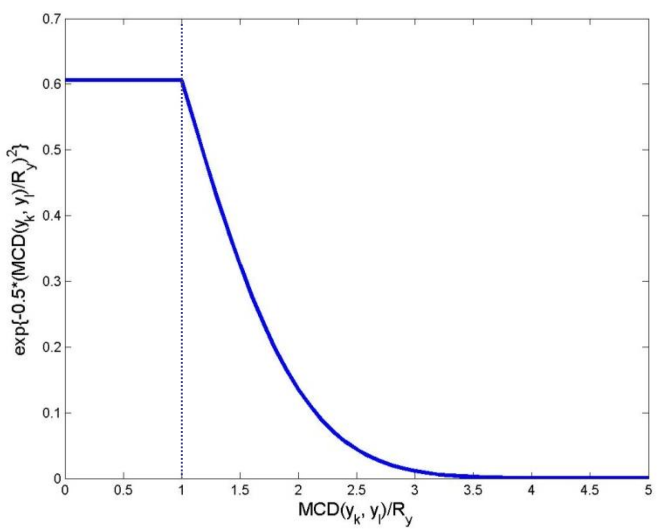
GB Conversion

Evidence Modeling

- A transition probability $\mathbf{y}^l \rightarrow \mathbf{y}^k$ at time t



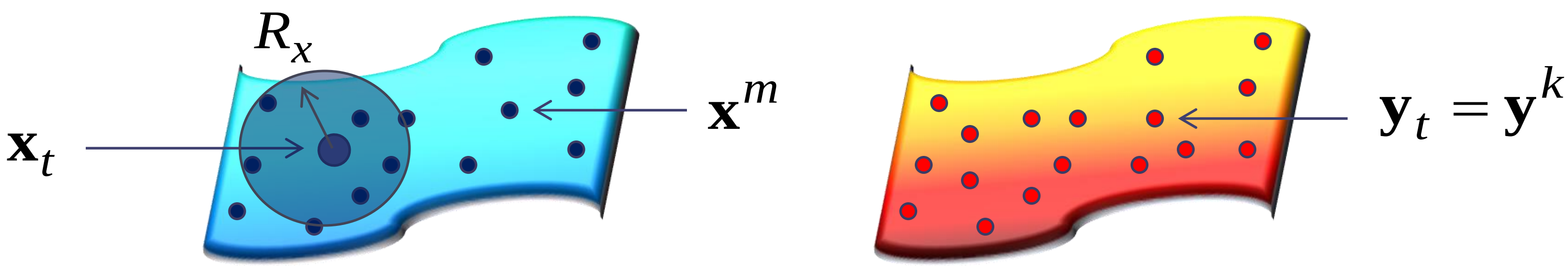
$$p\left(\mathbf{y}_t = \mathbf{y}^k \middle| \mathbf{y}_{t-1} = \mathbf{y}^l\right) \propto \exp\left\{-\frac{1}{2}\max\left(\frac{\mathrm{MCD}\left(\mathbf{y}^k, \mathbf{y}^l\right)}{R_y}, 1\right)^2\right\}$$



  - R_y – a parameter

GB Conversion

Likelihood Modeling



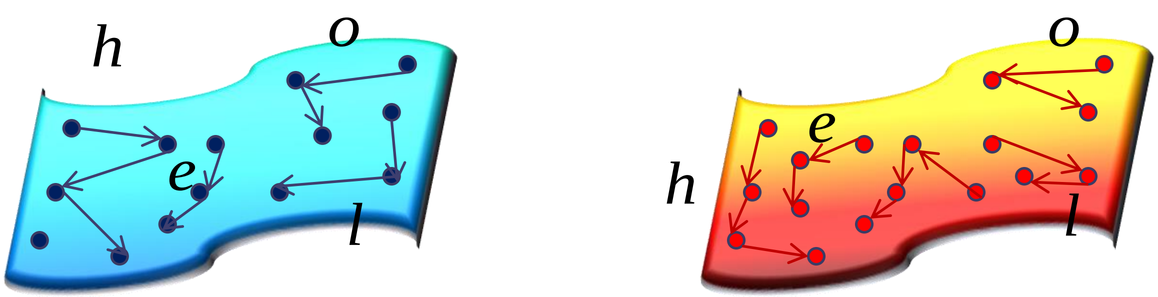
- We model the likelihood probability as:

$$p\left(\mathbf{x}_t \middle| \mathbf{y}_t = \mathbf{y}^k\right) \propto \sum_{m=1}^{N_x} p\left(\mathbf{x}^m \middle| \mathbf{y}_t = \mathbf{y}^k\right) \exp\left\{-\frac{1}{2}\left(\frac{\mathrm{MCD}\left(\mathbf{x}_t, \mathbf{x}^m\right)}{R_x}\right)^2\right\}$$

  - R_x – a parameter
  - $p\left(\mathbf{x}^m \middle| \mathbf{y}_t = \mathbf{y}^k\right)$ - the discrete likelihood

GB Conversion

Discrete Likelihood Modeling



- $p\left(\mathbf{x}^m \middle| \mathbf{y}_t = \mathbf{y}^k\right)$ - the correspondence between the source and target training vectors
- A parallel and phonetically labeled data

$$p\left(\mathbf{x}^m \middle| \mathbf{y}_t = \mathbf{y}^k\right) \propto \begin{cases} 1 & \mathbf{x}^m \text{ and } \mathbf{y}^k \text{ belong to the same phonetic sequence} \\ 0 & \text{otherwise} \end{cases}$$

GB Conversion Algorithm Summary

- **Input:** a sequence of source feature vectors $\mathbf{x}_{1:T}$
- Main Iteration: for t = 1, …T , perform the following steps:

1. Evaluate the prior weights: $w_{t|t-1}^{k}$
   - Using the evidence probability
2. Evaluate the posterior weights: $w_{t|t}^{k}$
   - Using the discrete and continuous likelihood probabilities
3. Obtain the converted spectra:

$$\hat{\mathbf{y}}_t = \sum_{k=1}^{N_y} w_{t|t}^{k} \mathbf{y}^k$$

- **Output:** a sequence of converted vectors: $\boxed{\hat{\mathbf{y}}_{1:T}}$

GB Conversion Algorithm Summary

- **Input:** a sequence of source feature vectors $\mathbf{x}_{1:T}$
- Main Iteration: for t = 1, …T , perform the following steps:

1. Evaluate the prior weights: $w_{t|t-1}^{k}$
   - Using the evidence probability
2. Evaluate the posterior weights: $w_{t|t}^{k}$
   - Using the discrete and continuous likelihood probabilities
3. Obtain the converted spectra:

$$\hat{\mathbf{y}}_t = \sum_{k=1}^{N_y} w_{t|t}^{k} \mathbf{y}^k$$

Offline

Online

- **Output:** a sequence of converted vectors: $\boxed{\hat{\mathbf{y}}_{1:T}}$

Experiments Results

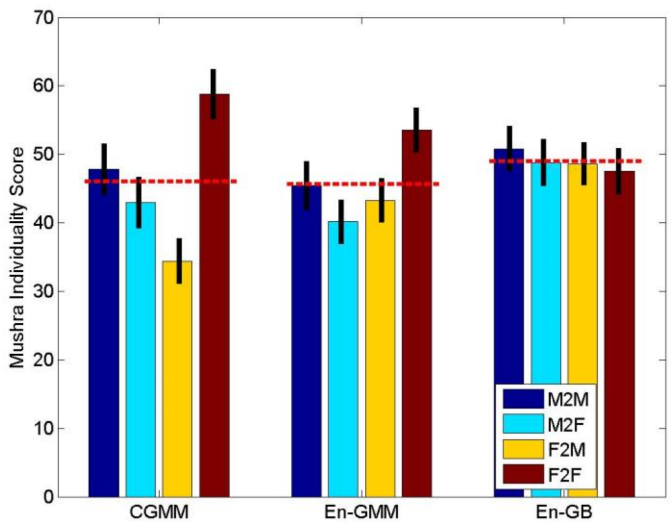
Objective Evaluations

- **Training set –** 10 parallel sentences
- **Testing set –** 50 sentences

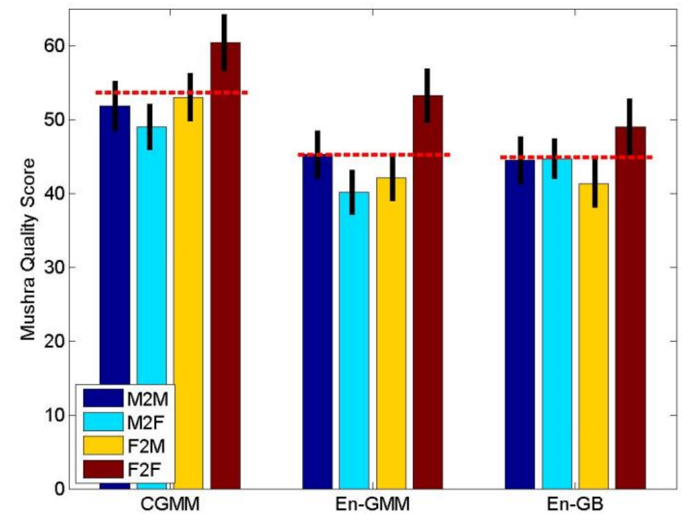
Gender	Conversion Method	ND	NGV
M2M	GMM + Modular Enhancement ϑ=2dB	0.74	0.55
	CGMM	0.82	0.45
	GB + Modular Enhancement ϑ=2dB	**0.73**	**0.6**
M2F	GMM + Modular Enhancement ϑ=2dB	0.74	0.54
	CGMM	0.84	0.46
	GB + Modular Enhancement ϑ=2dB	**0.73**	**0.68**
F2M	GMM + Modular Enhancement ϑ=2dB	0.75	0.69
	CGMM	0.85	0.61
	GB + Modular Enhancement ϑ=2dB	**0.77**	**1.1**
F2F	GMM + Modular Enhancement ϑ=2dB	**0.85**	0.65
	CGMM	0.89	0.6
	GB + Modular Enhancement ϑ=2dB	0.87	**0.98**

Subjective Evaluations

Individuality Tests



Quality Tests



Source	Target	GMM	CGMM	En-GMM	GB	EN-GB

Voice Conversion

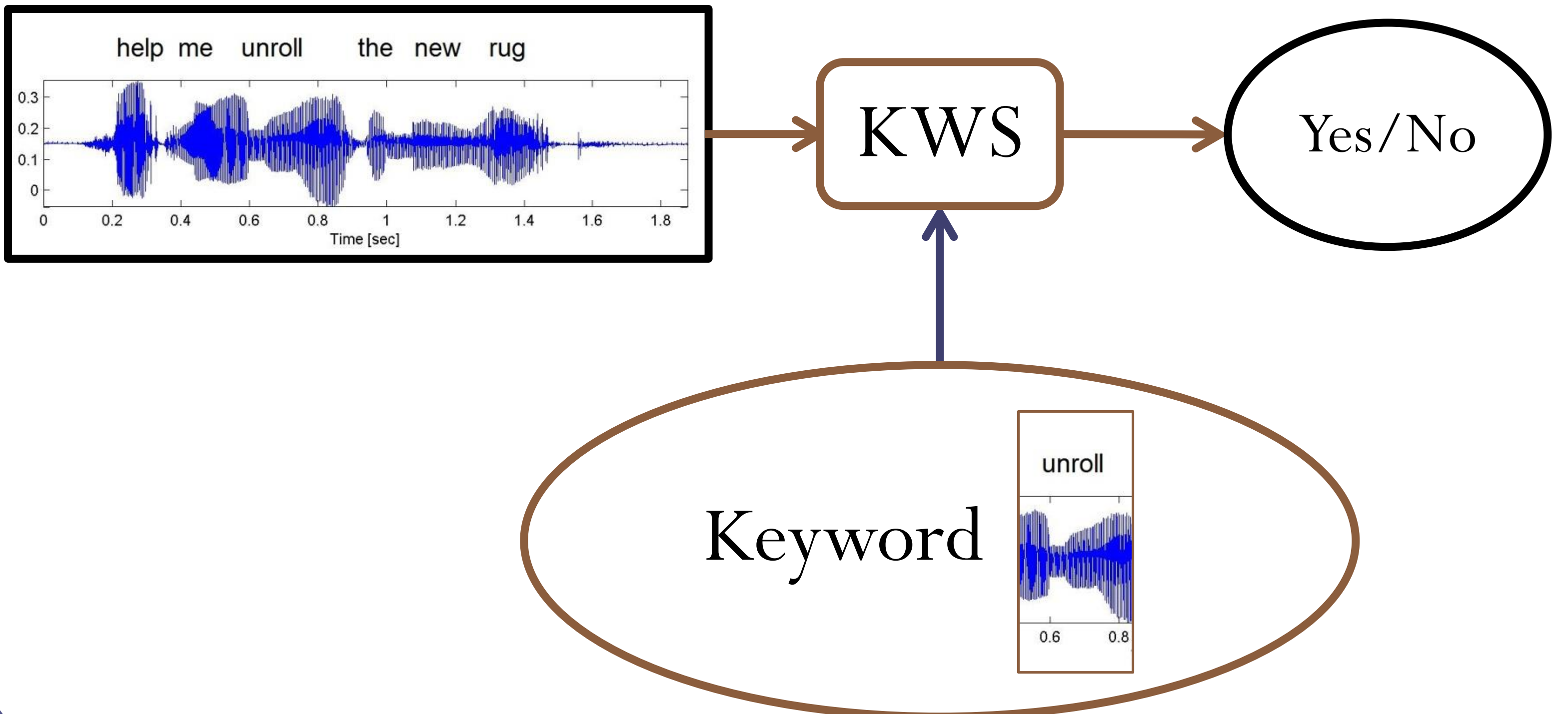
Keyword Spotting

Global Variance Enhancement

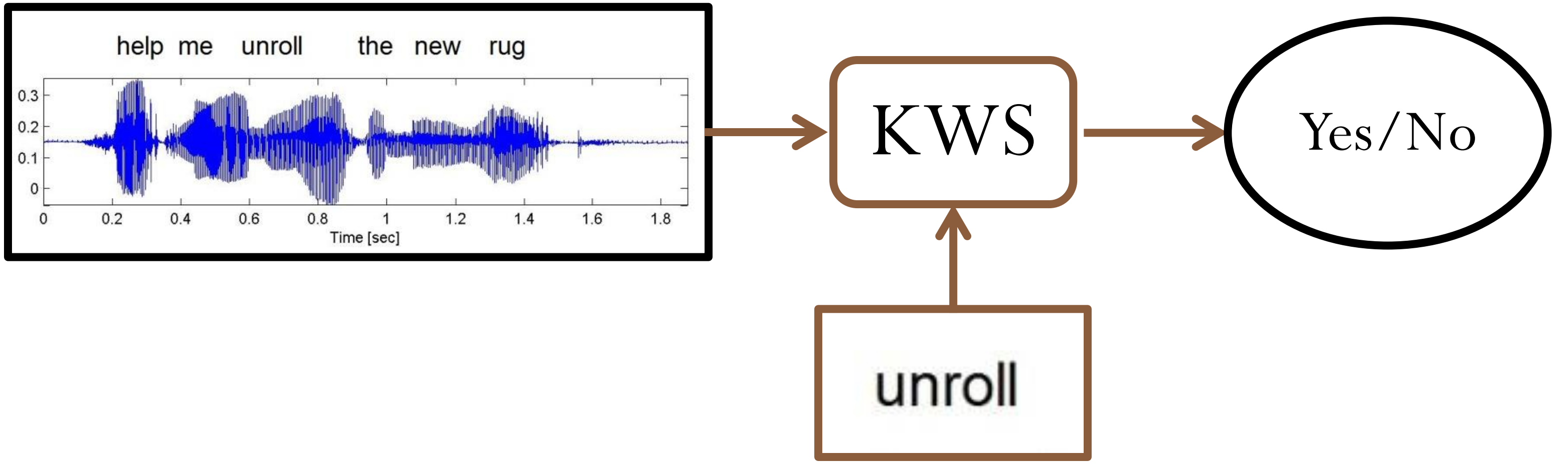
Grid-Based Conversion

Keyword Spotting (KWS)

- A task of detecting whether a keyword was said in a given speech utterance

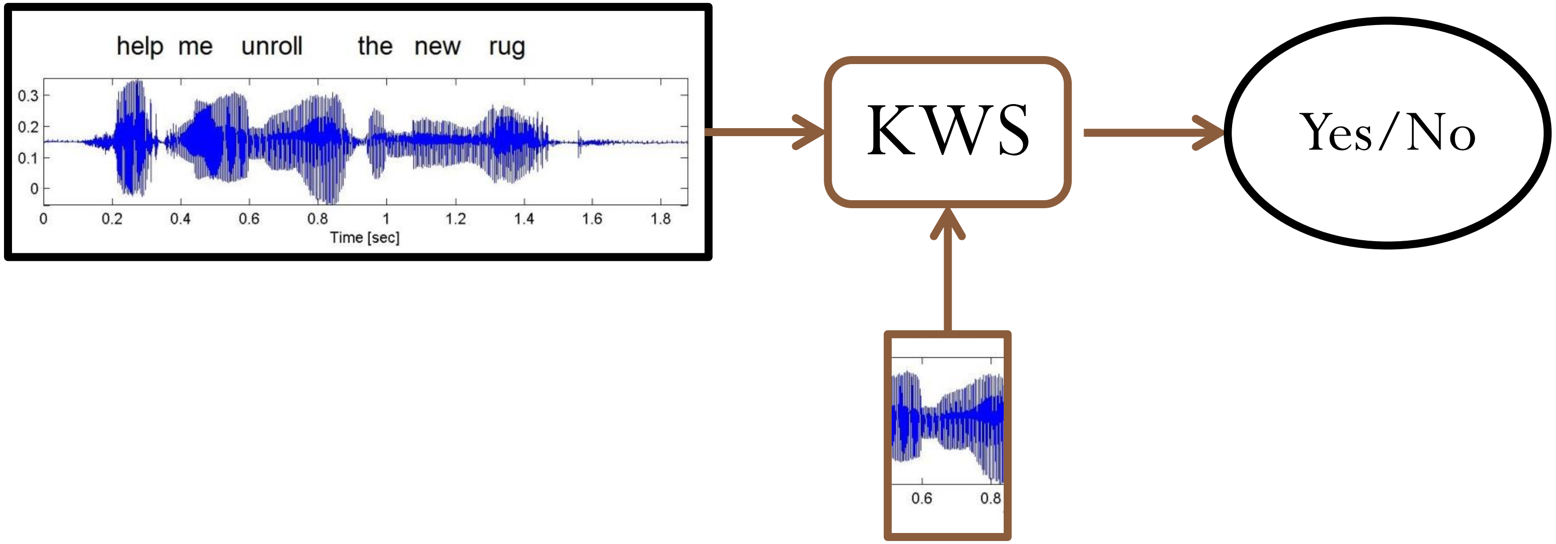


Input Keyword - Text



- Automatic Speech Recognition/ Phonetic Recognition
  - **Requires an enormous amount of annotated data and lexical resources**

Input Keyword – Speech Query By Example (QBE)



- Supervised methods
  - Use phonetically labeled recordings
- Unsupervised methods
  - Do not require any kind of labeled resource

Low Resource Environments

Computational limitations/ Recording time → Very small training set

Under documented languages/ speakers → No phonetic labeling

- Standard systems - based on HMM
  - Require medium-large data sets for training
  - Mostly require phonetic segmentation

Generative Vs. Discriminative

- **<u>Generative - HMM</u>**
  - Aim to statistically model the generation of the signal
  - Inference –
    - Using a likelihood score
  - **Does not directly maximize the detection rate**
- **<u>Discriminative</u>**
  - Usually based on a fixed length representation of speech utterances
  - **Training a classifier by maximizing the detection rate**
    - SVM, Perceptron, etc.

Common Discriminative Methods

Phonetic Segmentation

- New feature representation for speech utterances based on the estimated duration of phonemes and transition times [Keshet et al.,2009]
- **Requires phonetically segmented data (TIMIT – several hours)**

Common Discriminative Methods

Phonetic Segmentation

- New feature representation for speech utterances based on the estimated duration of phonemes and transition times [Keshet et al,2009]
- **Requires phonetically segmented data (TIMIT – several hours)**

Time-Frequency Representation

- A fixed length representation of the keyword based on:
  - Sprectro-temporal patches [Ezzat et al., 2008]
  - Patterns of high-energy tracks [Barnwal et al., 2012]
- **Use few positive examples and several minutes of negative speech, no metadata is needed**

Common Discriminative Methods

Phonetic Segmentation

- New feature representation for speech utterances based on the estimated duration of phonemes and transition times [Keshet et al,2009]
- **Requires phonetically segmented data (TIMIT – several hours)**

Time-Frequency Representation

- A fixed length representation of the keyword based on:
  - Sprectro-temporal patches [Ezzat et al., 2008]
  - Patterns of high-energy tracks [Barnwal et al., 2012]
- **Use few Positive Examples and several minutes of negative speech, no metadata is needed**

Proposed Approach

A Discriminative Classifier

Unsupervised

- No metadata is needed

Low Resources

- Trained using few positive examples and several minutes of negative speech

Proposed Concept: Training – Stage 1

- GMM - trained using unlabeled data
- Unlabeled data - easy to aquire according to the expected language/speakers

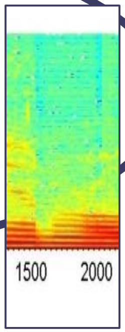
Bag-of-Gaussians

- Bag-of-Words
  - A known method in Natural Language Modeling (NLP)
  - Used for classification of documents (spam for example)
  - Sparse histograms - # occurrences of each word in a document
- Bag-of-Features
  - Used for image segmentation/classification
- Bag-of-Gaussians
  - A sparse histogram representing keyword

Histogram Representation For Keywords

1) Spectral features of a keyword

$$\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{T_w}\right)_{P\times T_w}$$



2) Posterior matrix

$$\begin{pmatrix} \vdots & \vdots & \vdots \\ \vdots & P\left(m \middle| \mathbf{x}_t; \alpha_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right) & \vdots \\ \vdots & \vdots & \vdots \end{pmatrix}_{M\times T_w}$$

3) Indicators

$$\mathbf{u} = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & 0 & \cdots & 0 \\ 0 & \vdots & \cdots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{M\times T_w}$$

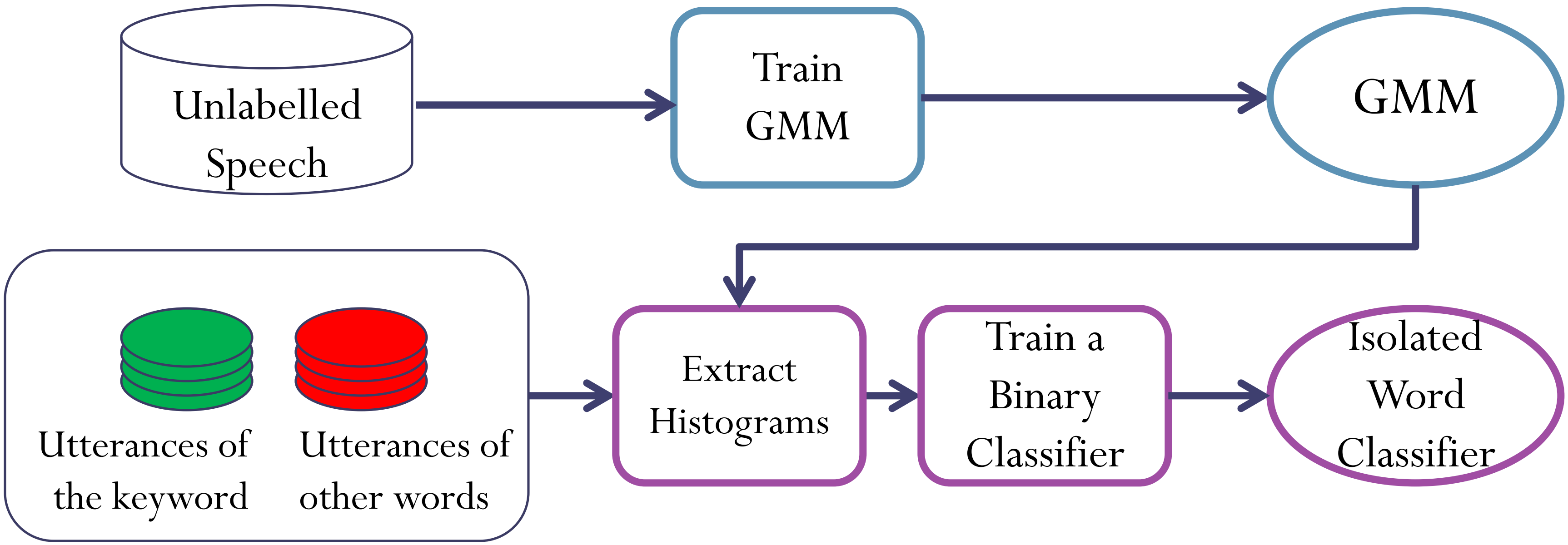
4) Histogram

$$\mathbf{v} = \frac{1}{T_w}\sum_{t=1}^{T_w}\mathbf{u}_t \in \Re^M$$

GMM Parameters

$$\left\{ \begin{matrix} \lambda^m, \mu^m, \boldsymbol{\Sigma}^m \\ m = 1, \ldots, M \end{matrix} \right\}$$

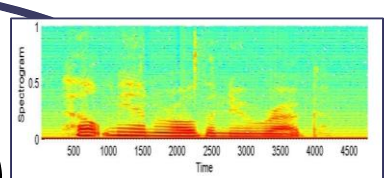
Proposed Concept: Training– Stage 2



Sentence Representation

Spectral features of a sentence

$$\left(\mathbf{x}_1,\ldots,\underbrace{\mathbf{x}_t,\ldots,\mathbf{x}_{t+T_w}}_{\mathbf{v}_t},\ldots,\mathbf{x}_{T_s}\right)_{P\times T_s}$$



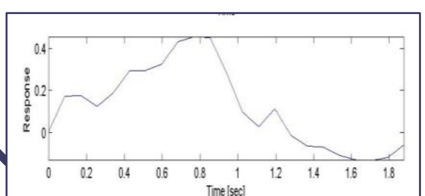
A sequence of histograms

$$\left(\mathbf{v}_1,\ldots,\mathbf{v}_t,\ldots,\mathbf{v}_{\tau_s}\right)_{M\times\tau_s}$$

Isolated Word Classifier

Response Curve

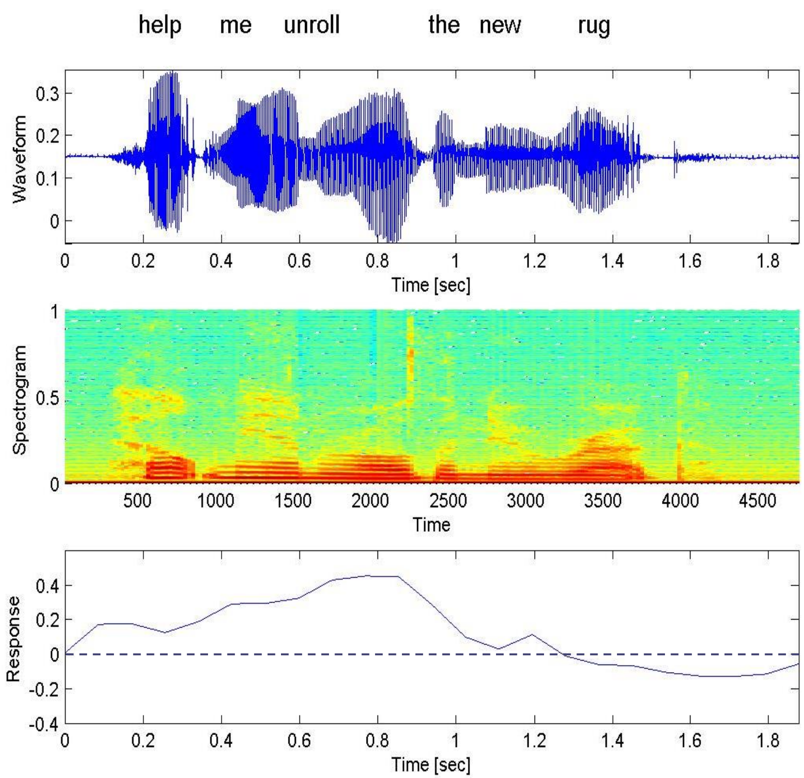
$$\mathbf{S}_{1:\tau_s}=\left(S_1,\ldots,S_{\tau_s}\right)\in\Re^{\tau_s}$$



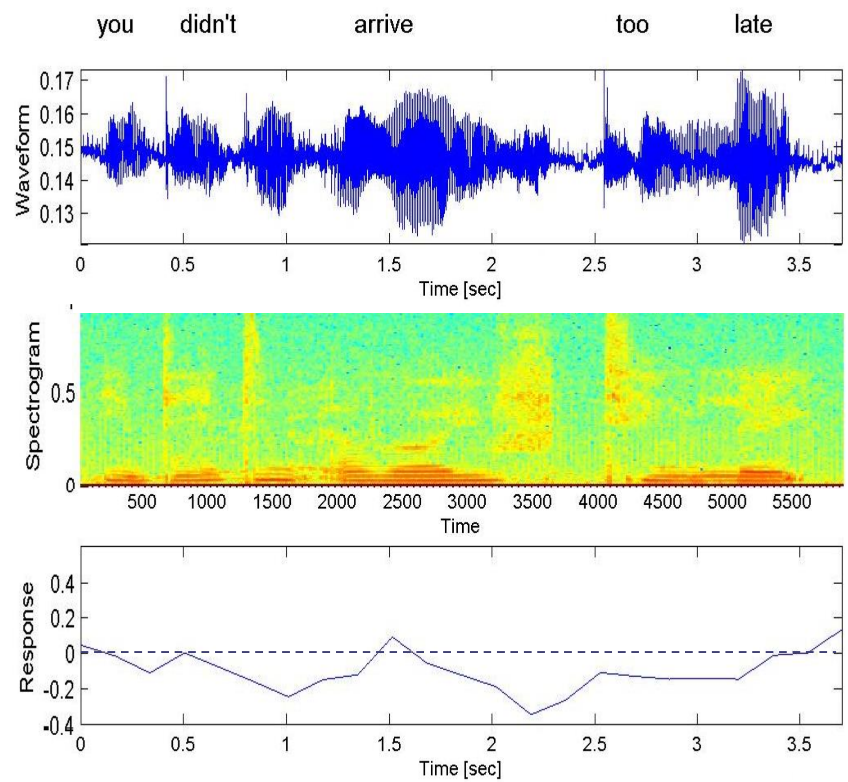
Global Features ϕ

Sentence Representation – Cont'd

A Positive Sentence



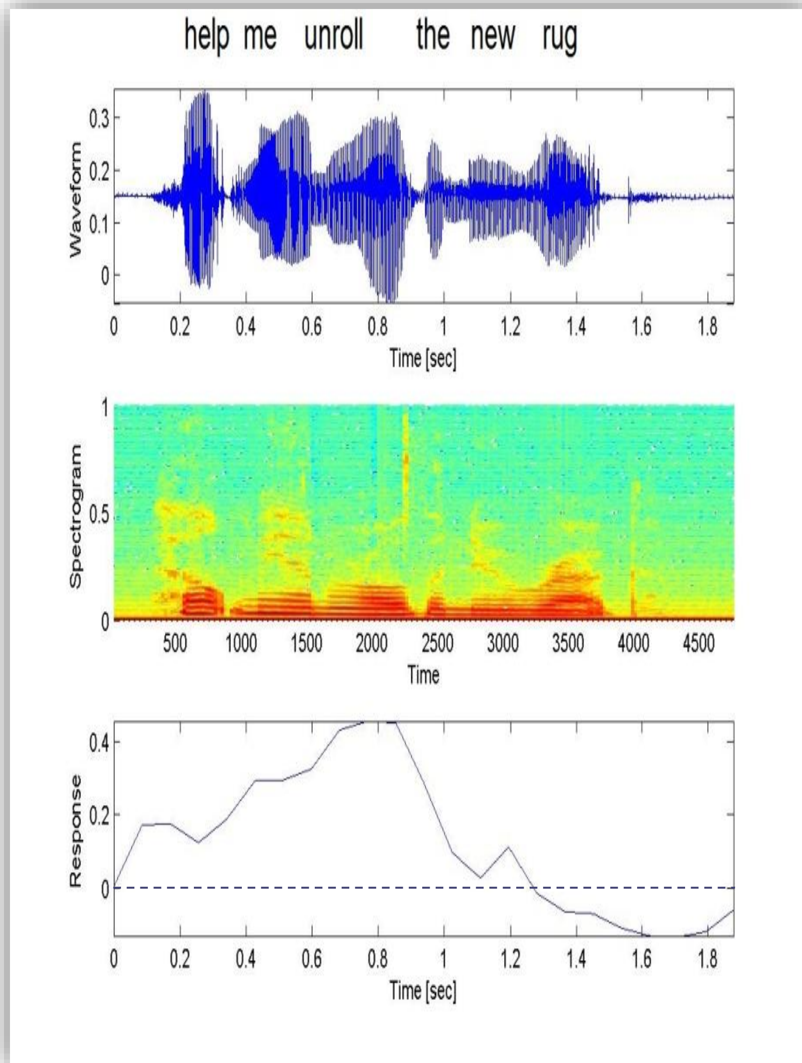
A Negative Sentence



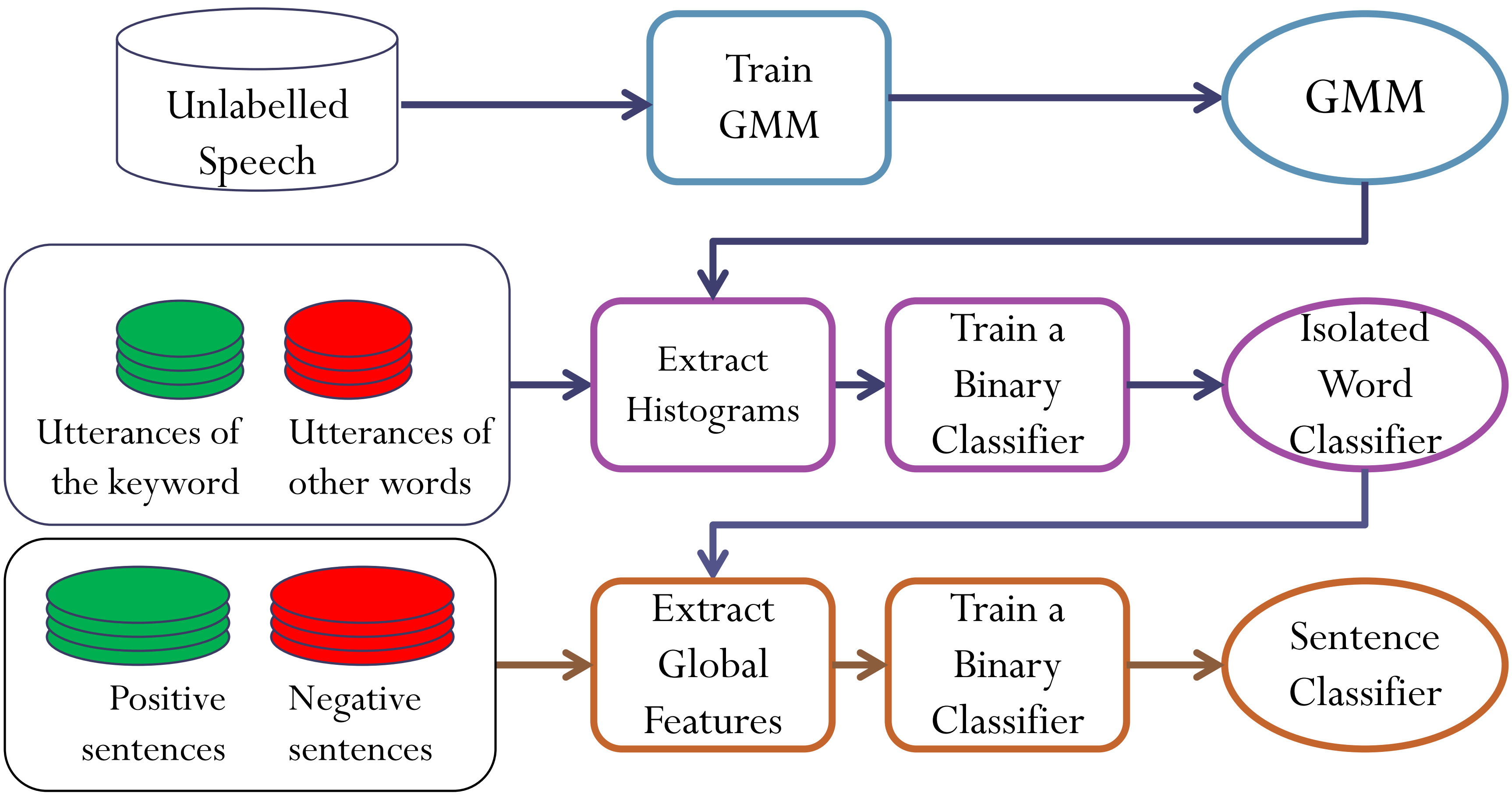
Sentence Representation – Cont'd

- Instead of using a threshold we generalize:
  - Train a binary classifier using the following features extracted from the response curve:
- Where: $\phi = \left(M_x, m_n, a, DR, \delta, \delta^2\right)$
  - M_x - maximum value*
  - m_n - minimum value*
  - a - mean value*
  - DR - dynamic range*
  - δ - mean first derivative
  - δ^2 - mean second derivative

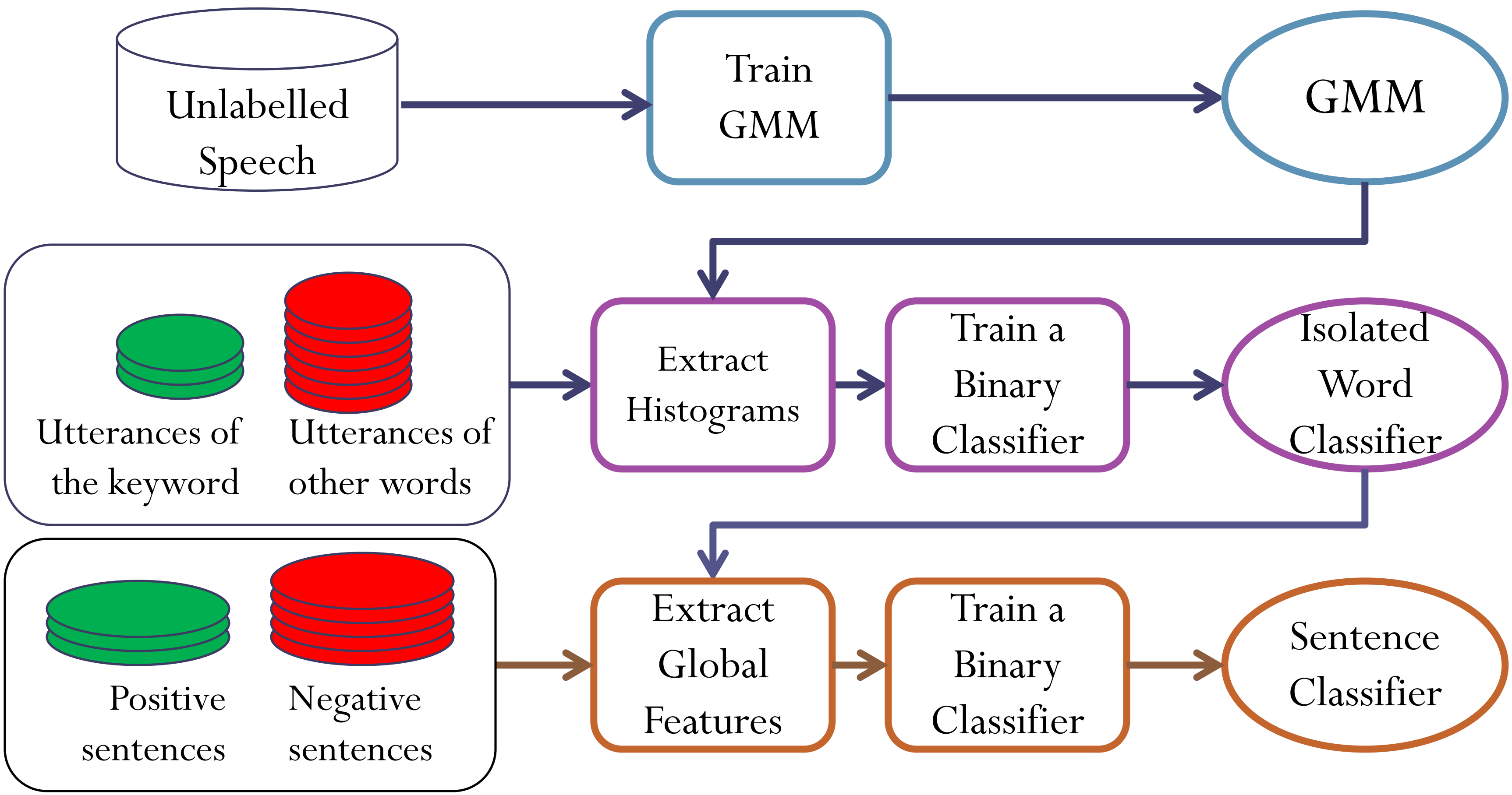
*Normalized by the std



Proposed Concept: Training– Stage 3



Unbalanced Training Set

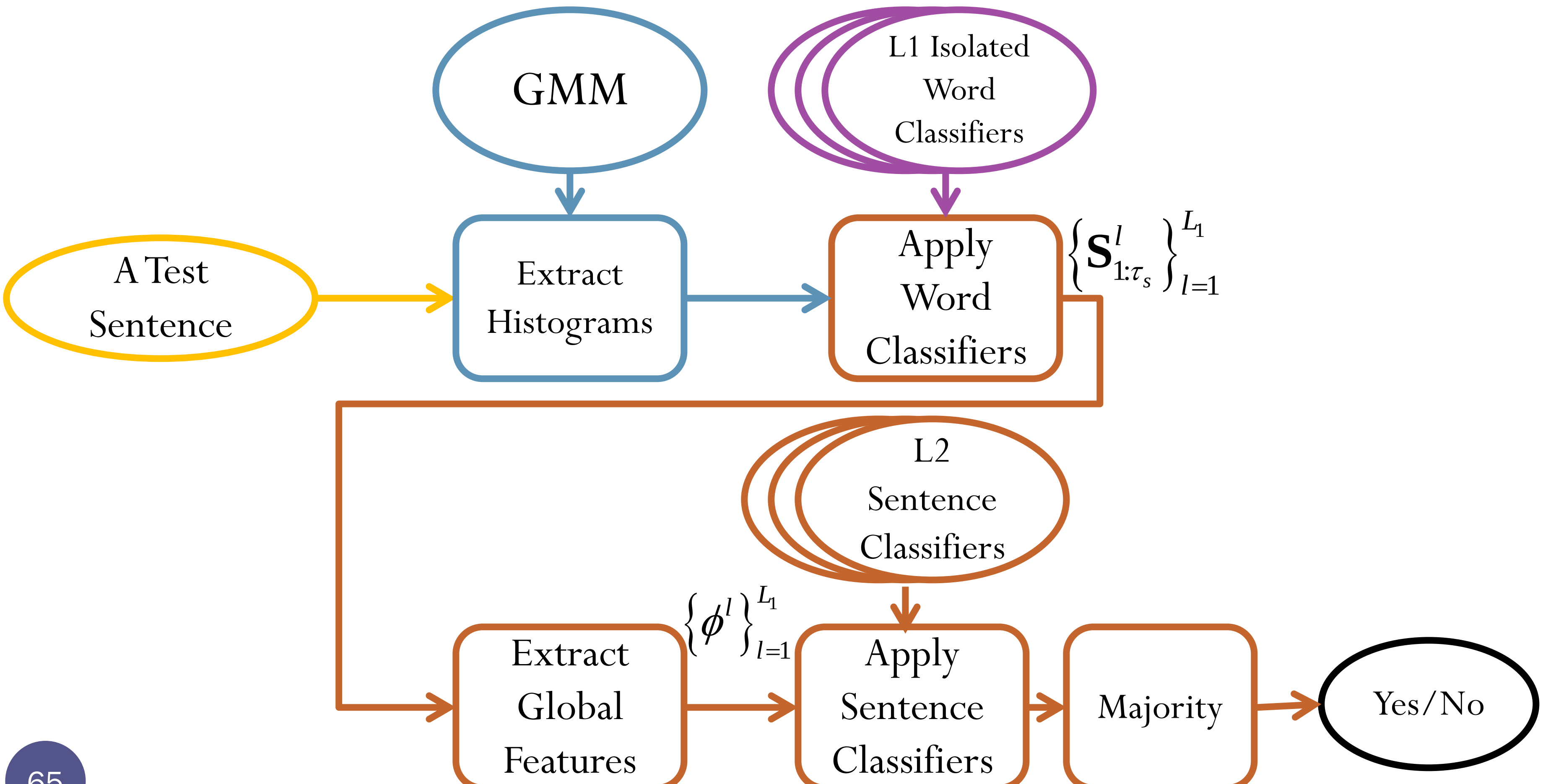


Bagging Predictors [Breiman, 1996]

- Labeled samples - harder to acquire
- Positive Examples << Negative Examples
- Training using all negative data:
  - Increase robustness
  - A biased classifier
- **Bagging predictors - having the best of both:**
  - Uniformly sample L subsets of negative examples
  - Train L binary classifiers
- **Inference – apply all classifiers and take the majority decision**

Proposed System

- We use bagging predictors for isolated word classification and for sentence classification:



Experiments Results

Experiments

- **Adults speech (TIMIT)**
  - Following a previously presented protocol [Ezzat et al, 2008; Barnwal et al., 2012]:
    - Amount of positive examples - five set sizes - 5,10,50,100 and 200
    - Amount of negative examples - constant size - 100 sentences
- **Children's speech (CSLU)**
  - Clean speech
  - Noisy speech
    - Babble and car -5dB to 20dB

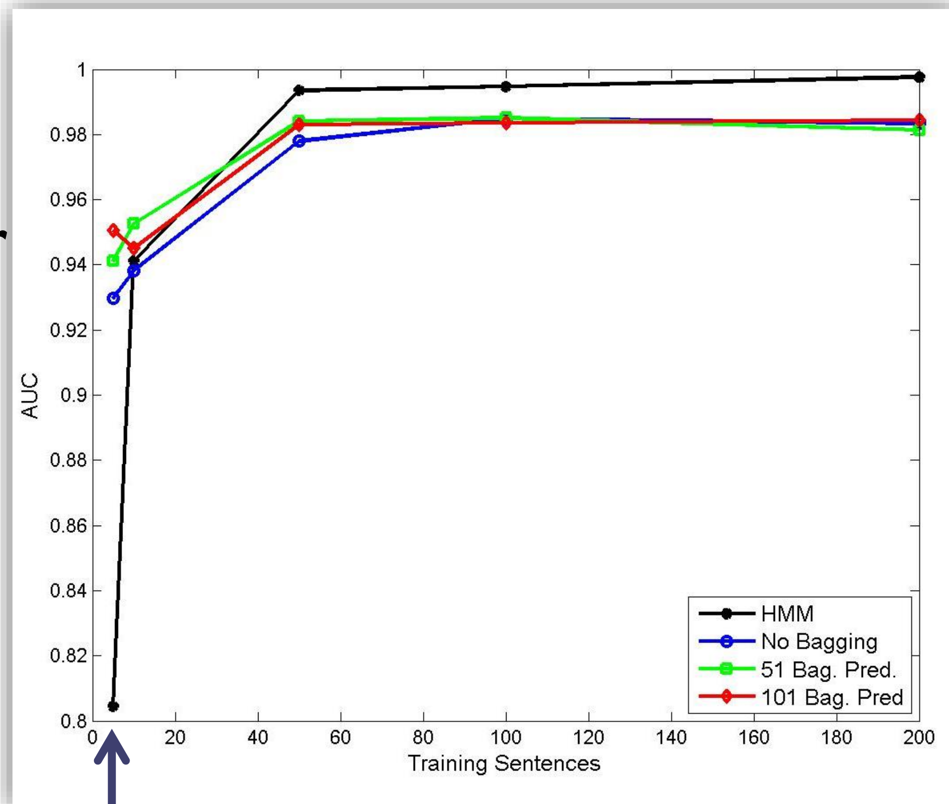
TIMIT Experiments (Adults)

Detection of Four Words: "greasy", "dark", "wash", and "oily"

- AUC – Area under the curve
- Averaged over detection of four words
  - HMM
  - Proposed system

➢ **The proposed system is better for 5-10 positive examples**

➢ **Bagging is more substantial for small training sets**

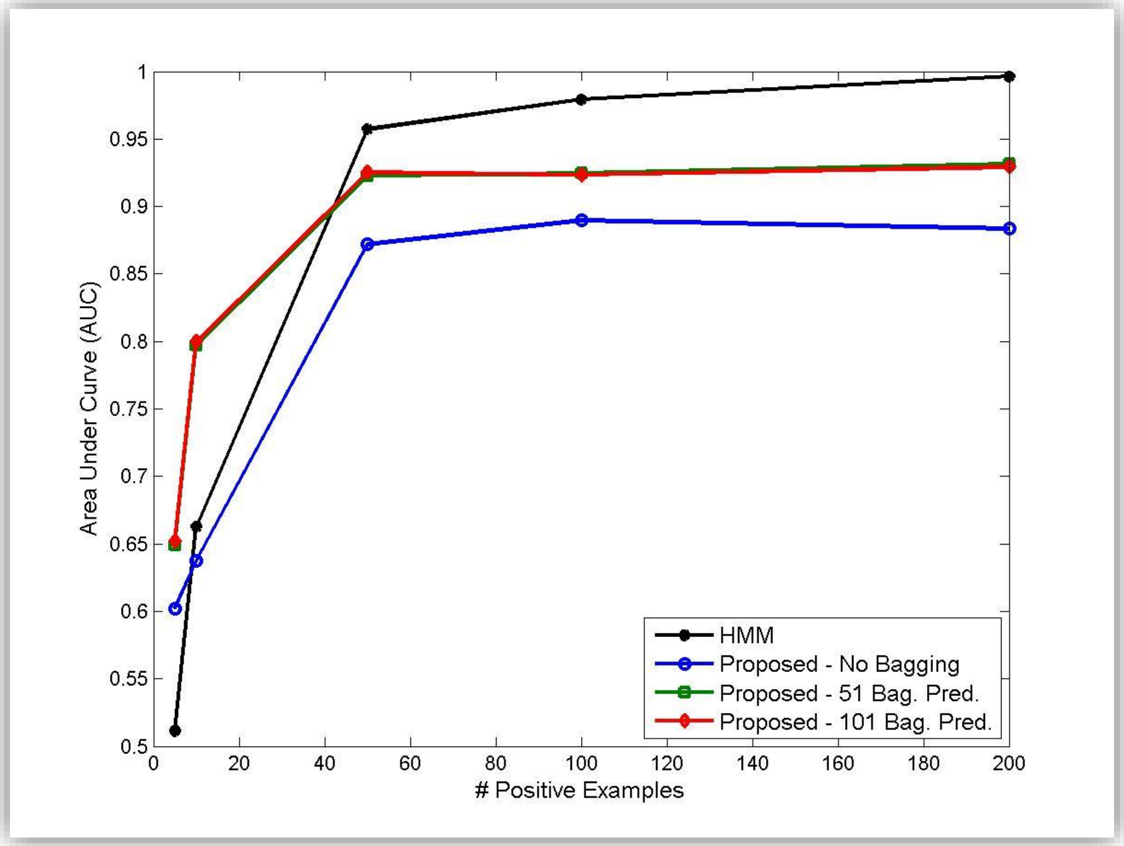


5 Positive examples

CSLU Experiments - Children's Speech

Detection of Three Words: "one", "two", "unroll"

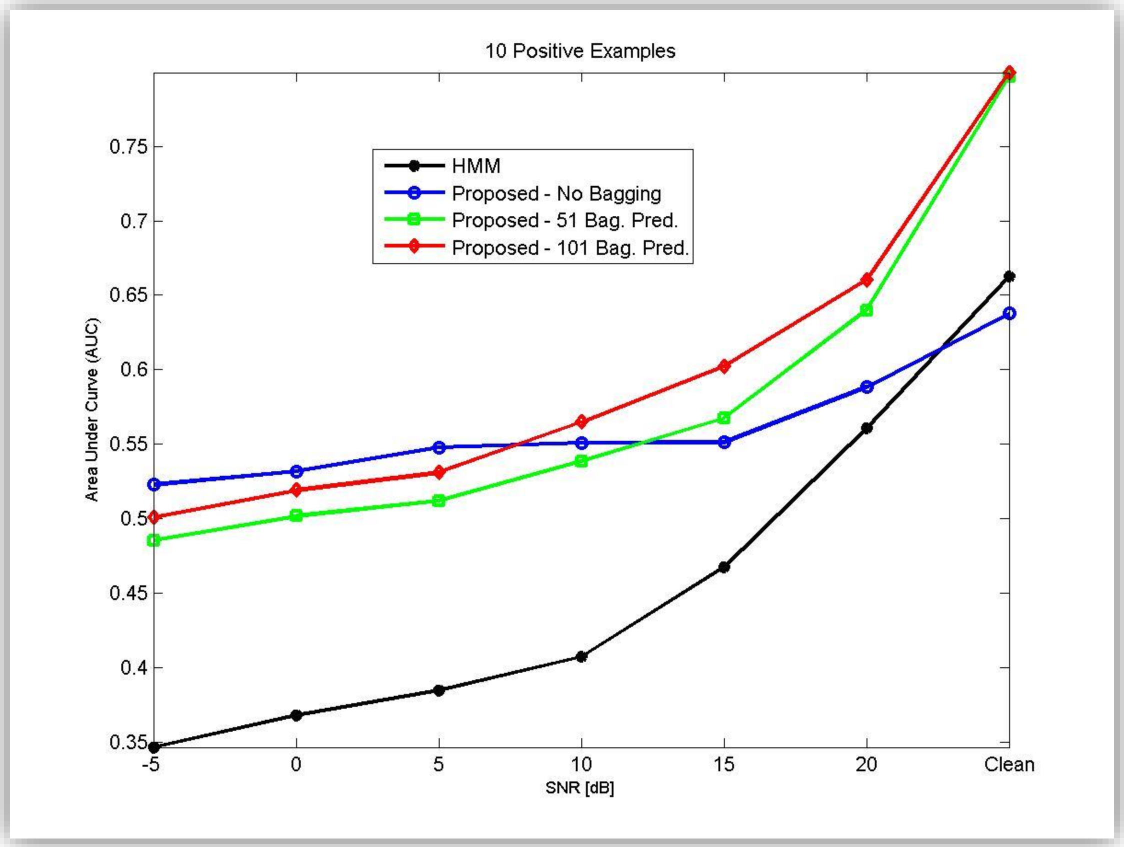
- Age – kindergarten-5th grade
- Training – clean signals
- Testing – clean signals



CSLU Experiments - Children's Speech

Detection of Three Words: "one", "two", "unroll"

- Age – kindergarten-5th grade
- Training – clean signals, 10 positive examples
- Testing – noisy signals (babble)



Summary - Main Contributions - 1

<u>Voice Conversion</u>

- **Global Variance Enhancement:**
  I. Embedded in GMM training (CGMM)
  II. Modular post processing block
- **Grid-Based Conversion**
  - Sequential estimation using Bayesian tracking

Improved Speech Quality

Low Resource Applications

Summary - Main Contributions - 2

Keyword Spotting:

- Discriminative
- Unsupervised
- **A histogram representation for keywords**
- **Global features representation for sentences**
- **Bagging predictors**

Low resource applications

Robust to:

- Training data size
- Children's speech
- Noise

Future Work

- **Voice Conversion**
  - Modeling and conversion of prosody features: pitch, duration and energy
  - Alternative measures for objective evaluation with better correspondence to subjective results
- **Keyword Spotting**
  - Histogram representation of keywords – alternative modeling considering the temporal context of spectral feature vectors
  - Global features representation of sentences – explore new features for improved representation and classification of positive and negative response curves

Thank You