ICASSP 2008 Survey

Presenter: Shih-Hsiang Lin

Outline

- MMSE-BASED STEREO FEATURE STOCHASTIC MAPPING FOR NOISE ROBUST SPEECH RECOGNITION (Oral)
  - Xiaodong Cui, IBM T. J. Watson Research Center, United States; Mohamed Afify, MSA university, Egypt; Yuqing Gao, IBM T. J. Watson Research Center, United States
- MINIMUM BAYES-RISK DECODING WITH PRESUMED WORD SIGNIFICANCE FOR SPEECH BASED INFORMATION RETRIEVAL (Poster)
  - Takashi Shichiri, Hiroaki Nanjo, Takehiko Yoshimi, Ryukoku University, Japan
- CRANDEM SYSTEMS: CONDITIONAL RANDOM FIELD ACOUSTIC MODELS FOR HIDDEN MARKOV MODELS (Oral)
  - Eric Fosler-Lussier, Jeremy Morris, The Ohio State University, United States
- DISCRIMINATIVE FEATURE WEIGHTING USING MCE TRAINING FOR TOPIC IDENTIFICATION OF SPOKEN AUDIO RECORDINGS (Oral)
  - Timothy Hazen, Anna Margolis, MIT Lincoln Laboratory, United States
- BROADCAST NEWS SUBTITLING SYSTEM IN PORTUGUESE (Poster)
  - João Neto, INESC ID / IST, Portugal; Hugo Meinedo, Márcio Viveiros, Renato Cassaca, Ciro Martins, INESC ID, Portugal; Diamantino Caseiro, INESC ID / IST, Portugal

MMSE-BASED STEREO FEATURE STOCHASTIC MAPPING FOR NOISE ROBUST SPEECH RECOGNITION

Xiaodong Cui[1], *Mohamed Afify*[2] *and Yuqing Gao*[1]

IBM T. J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY, 10598, USA[1]
ITIDA, Ministry of Communications and Information Technology, Cairo, Egypt[2]
Emails: cuix@us.ibm.com, mohamed_afify2001@yahoo.com, yuqing@us.ibm.com

Introduction

- The approaches of using stereo data are able to learn the statistical relationship between clean and noisy speech signals directly from the data for denoising
  - requiring no model between clean and noisy speech signals
- In their previous work, they proposed an iterative MAP-based stochastic mapping approach utilizing stereo data
  - a GMM distribution is assumed for the joint stereo features
  - he estimation of the clean feature from the noisy feature was carried out iteratively by the EM algorithm
- In this paper, they propose an MMSE estimate of the clean feature is derived which can be shown as a piece-wise linear function

MMSE Mathematical Formulation

- Assume we have a set of stereo data $\{(x_i, y_i)\}$
- Define $z \equiv (x, y)$ as the concatenation of the two channels
- The first step in constructing the mapping is training the joint probability model for *p(z)*

$$p(z) = \sum_{k=1}^{K} c_k N(z; \mu_{z,k}, \Sigma_{zz,k}) \quad \text{where} \quad \mu_{z,k} = \begin{pmatrix} \mu_{x,k} \\ \mu_{y,k} \end{pmatrix} \quad \Sigma_{z,k} = \begin{pmatrix} \Sigma_{xx,k} & \Sigma_{xy,k} \\ \Sigma_{yx,k} & \Sigma_{yy,k} \end{pmatrix}$$

- Given the observed noisy speech feature *y*, the MMSE estimate of clean speech *x* is given by

$$\begin{aligned} \hat{x} &= E[x|y] \\ &= \int_x p(x|y) x dx \\ &= \sum_k p(k|y) \int_x p(x|k,y) x dx \\ &= \sum_k p(k|y) E[x|k,y] \end{aligned}$$

where

$$\begin{aligned} E[x|k,y] &= \mu_{x|y,k} \\ &= \mu_{x,k} + \Sigma_{xy,k} \Sigma_{yy,k}^{-1} (y - \mu_{y,k}) \end{aligned}$$

MMSE Mathematical Formulation (cont.)

- It is obvious that the MMSE estimate of x is a piece-wise linear function of the noisy feature y, as we can re-write in the following form

$$\hat{x} = \sum_k p(k|y)(A_k y + b_k)$$

$$\begin{aligned} A_k &= \Sigma_{xy,k}\Sigma_{yy,k}^{-1} \\ b_k &= \mu_{x,k} - \Sigma_{xy,k}\Sigma_{yy,k}^{-1}\mu_{y,k} \end{aligned}$$

MMSE vs. SPLICE

- In SPLICE, the estimate of clean feature is obtained as

$$\hat{x} = \sum_k p(k|y)(y + r_k)$$

  - where the bias is estimated by utilizing stereo-data

$$r_k = \frac{\sum_n p(k|y_n)(x_n - y_n)}{\sum_n p(k|y_n)}$$

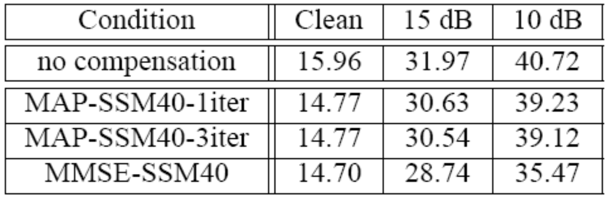
- Comparison
  - The posterior probability in SPLICE is computed from the noisy feature distribution while MMSE is computed from the joint distribution
  - SPLICE assumes the transformation matrix is an identity matrix, which is a special case of the MMSE when $\Sigma_{xy,k} = \Sigma_{yy,k}$
  - If a perfect correlation is assumed between the clean feature and noisy feature, then $p(k|x_n)$ and $p(k|y_n)$ are approximately identical from the joint GMM distribution

    ➔ $r_k = \mu_{x,k} - \mu_{y,k}$

Experimental Results

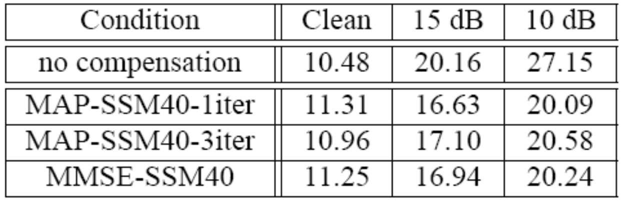
- Experiments are performed on large vocabulary spontaneous speech recognition system
  - Both clean and multi-style (MST) acoustic models are trained and tested
    - There are in total about 120 hours of clean data in the training set
    - In the MST model case, 15dB and 10dB noisy data are generated by adding humvee, tank and babble noise to the clean data
  - The experiments are carried out on two test sets both of which are collected in the DARPA Transtac project
    - The first test set (Set A) has 11 male speakers and 2070 utterances in total recorded in the clean condition.
      - The utterances are spontaneous speech which are corrupted artificially by adding humvee, tank and babble noise to produce 15dB and 10dB noisy test data
    - The second test set (Set B) has 7 male speakers with 203 utterances from each
      - The utterances were recorded in the real-world environment with humvee and tank noise running in the background
      - a very noisy evaluation set and utterance SNRs are measured around 5dB to 8dB.

Experimental Results (cont.)



Condition	Clean	15 dB	10 dB
no compensation	15.96	31.97	40.72
MAP-SSM40-1iter	14.77	30.63	39.23
MAP-SSM40-3iter	14.77	30.54	39.12
MMSE-SSM40	14.70	28.74	35.47

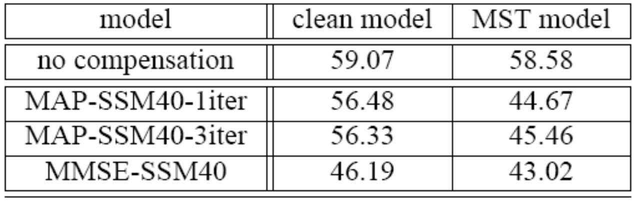
Table 2. Word error rate (WER) with clean acoustic model on Set A using MAP and MMSE mappings.



Condition	Clean	15 dB	10 dB
no compensation	10.48	20.16	27.15
MAP-SSM40-1iter	11.31	16.63	20.09
MAP-SSM40-3iter	10.96	17.10	20.58
MMSE-SSM40	11.25	16.94	20.24

Table 3. Word error rate (WER) with MST model on Set A using MAP and MMSE mappings.

- With clean acoustic model, the MAP mapping with 3 iterations obtains better performance than 1 iteration
- The MMSE mapping gives better performance than the MAP with 3 iterations
- When multi-style training is performed, both MAP MST and MMSE MST yield significant better performance compared to MST without noise compensation in 15dB and 10dB.



model	clean model	MST model
no compensation	59.07	58.58
MAP-SSM40-1iter	56.48	44.67
MAP-SSM40-3iter	56.33	45.46
MMSE-SSM40	46.19	43.02

Table 4. Word error rate (WER) with clean and MST model on Set B using MAP and MMSE mapping.

- In this real-world noisy test set, the MMSE mapping achieves 18% relative WER reduction compared to the MAP mappings in the clean model scenario

MINIMUM BAYES-RISK DECODING WITH PRESUMED WORD SIGNIFICANCE FOR SPEECH BASED INFORMATION RETRIEVAL

Takashi Shichiri, Hiroaki Nanjo, Takehiko Yoshimi

Graduate School of Science and Technology, Ryukoku University
Seta, Otsu 520-2194, Japan
{shichiri, nanjo, yoshimi}@nlp.i.ryukoku.ac.jp

Introduction

- Since the significance of words differs in IR, in ASR for IR,
  - ASR performance should be evaluated based on weighted word error rate (WWER)
    - gives a different weight on each word recognition error from the viewpoint of IR, instead of word error rate (WER)
    - words that greatly affect IR performance must be detected with higher priority

Correct : 請 幫 我 找 師 範 大 學 的 新 聞
ASR 1 : 請 幫 我 找 吃 飯 大 學 的 新 聞
ASR 2 : 請 綁 我 照 師 範 大 學 的 心 文

- Ideal weights would give a WWER equivalent to IR performance degradation when a corresponding ASR result is used as a query for the IR system

Evaluation Measure of ASR

- Word Error Rate (WER)

$$\mathrm{WER} = (I + D + S)/N$$

  - ***N*** is the of words in the correct transcript, ***I*** is the number of inserted words, ***D*** is the number of deleted words, ***S*** is the number of substituted words
  - all words are treated uniformly or with the same weight
  - However, there must be a difference in the weight of errors
    - since several keywords have more impact on IR or the understanding of the speech than trivial functional words

- Weighted Word Error Rate (WWER)

$$\begin{aligned} \mathrm{WWER} &= \frac{V_I + V_D + V_S}{V_N} \\ V_N &= \Sigma_{w_i} \quad v_{w_i} \\ V_I &= \Sigma_{\hat{w}_i \in I} \quad v_{\hat{w}_i} \\ V_D &= \Sigma_{w_i \in D} \quad v_{w_i} \\ V_S &= \Sigma_{seg_j \in S} \quad v_{seg_j} \\ & \quad v_{seg_j} = \max(\Sigma_{\hat{w}_i \in seg_j} v_{\hat{w}_i}, \Sigma_{w_i \in seg_j} v_{w_i}) \end{aligned}$$

WWER equals WER if all word weights are set to 1

Minimum Bayes-Risk Decoding

- Decoding strategy : Minimize WWER based on the Minimum Bayes-Risk framework

$$\delta(X) = \operatorname*{argmin}_{W} \sum_{W'} l(W, W') \cdot P(W'|X)$$

loss function

  - In order to minimize WER, Levenshtein distance or WER is used as a loss function
  - In this paper, they use WWER as the loss function

Information Retrieval – WEB Page Retrieval

- Retrieval using Word Statistics
  - The similarity between a query and documents is defined by the inner product of the feature vectors of the query and the specific document
    - TF-IDF is used as the feature vector

$$\text{TF-IDF}(t,i) = \frac{tf_{t,i}}{\frac{\text{DL}_i}{\text{avglen}} + tf_{t,i}} \cdot \log \frac{N}{df_t}$$

  - normalize TF values using length of the document (DL$_i$) and average document lengths over all documents (avglen) because longer document have more words and TF values tend to be larger

- Task
  - Web retrieval task distributed by NTCIR (NTCIR-3 WEB task)
  - For speech-based information retrieval, 470 query utterances by 10 speakers are also included

Information Retrieval – WEB Page Retrieval (cont.)

- Evaluation Measure of IR
  - For an evaluation measure of IR, discount cumulative gain (DCG) is used

$$\mathrm{DCG}(i) = \begin{cases} g(1) & \text{if } i = 1 \\ \mathrm{DCG}(i-1) + \dfrac{g(i)}{\log(i)} & otherwise \end{cases}$$

$$g(i) = \begin{cases} h & \text{if } d_i \in H \\ a & \text{else if } d_i \in A \\ b & \text{else if } d_i \in B \end{cases} \qquad \begin{matrix} \text{Highly relevant} \\ \text{Relevant} \\ \text{Partially relevant} \end{matrix}$$

    - d_i represents *i*-th retrieval result (document)
    - H, A, and B represent a degree of relevance
    - When retrieved documents include many relevant documents that are ranked higher, the DCG score increases
- For an evaluation measure of IR performance degradation, IR score degradation ratio (IRDR) is defined as below

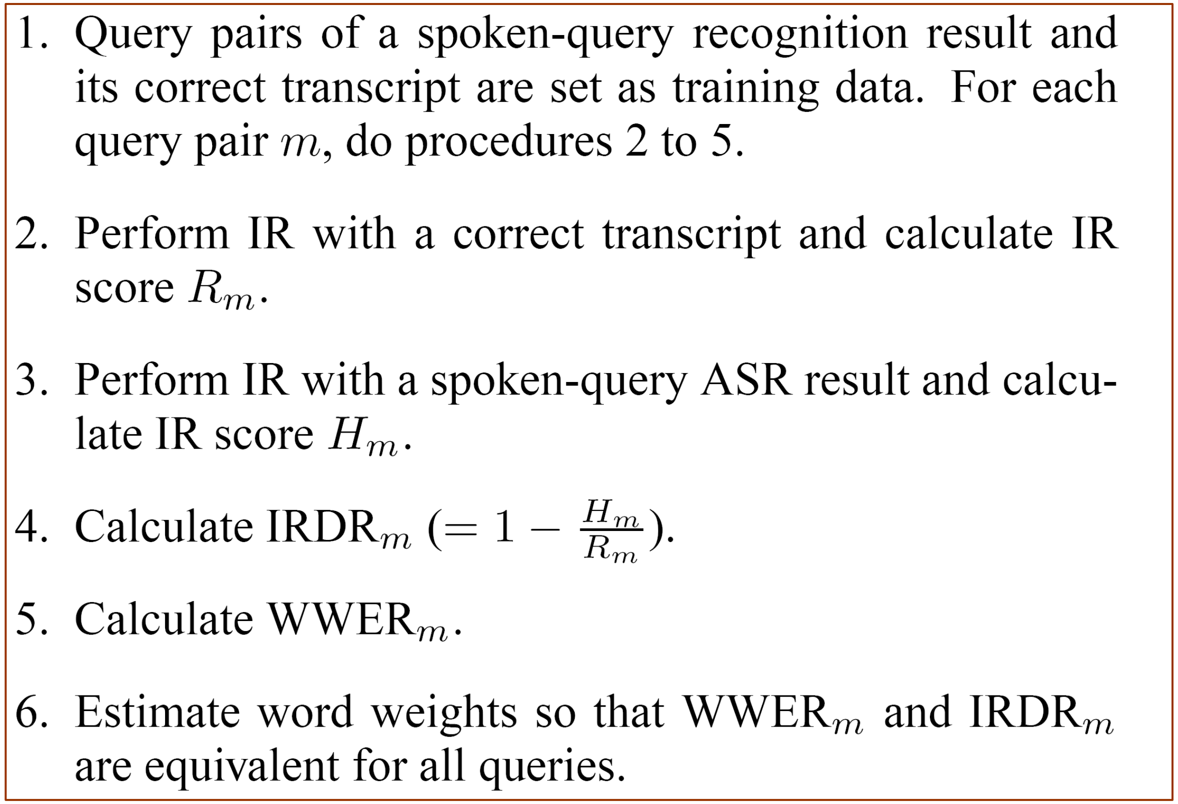
$$\mathrm{IRDR} = 1 - \frac{H}{R}$$

H represents a DCG score given by the ASR result of the spoken query

R represents a DCG score calculated with IR results by text query

Estimation of Word Weights

- A word weight should be defined based on its influence on IR
  - Specifically, weights are estimated so that WWER will be equivalent to an IR performance degradation (IRDR)



1. Query pairs of a spoken-query recognition result and its correct transcript are set as training data. For each query pair m, do procedures 2 to 5.
2. Perform IR with a correct transcript and calculate IR score R_m.
3. Perform IR with a spoken-query ASR result and calculate IR score H_m.
4. Calculate IRDR$_m$ ($= 1 - \frac{H_m}{R_m}$).
5. Calculate WWER$_m$.
6. Estimate word weights so that WWER$_m$ and IRDR$_m$ are equivalent for all queries.

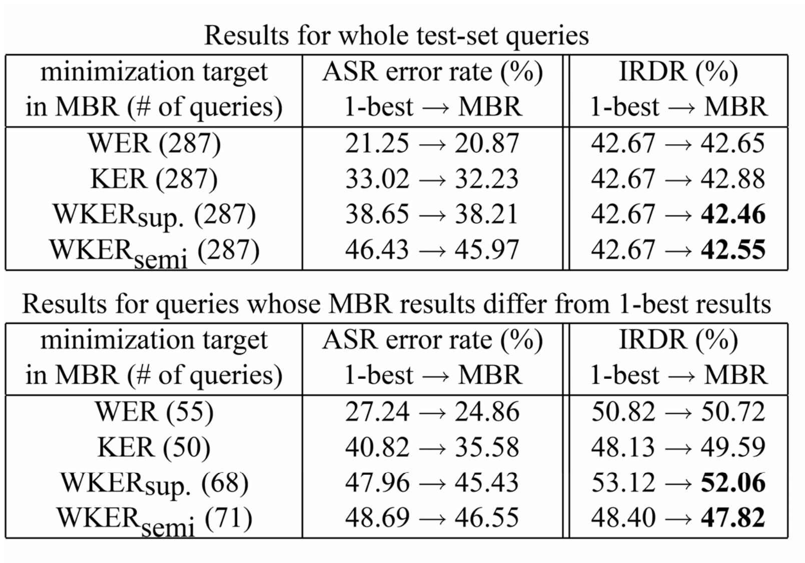
Estimation of Word Weights (cont.)

- Practically, procedure 6 is defined to minimize the mean square error between both evaluation measures (WWER and IRDR)

$$F(\mathbf{x}) = \sum_{m} \left(\frac{E_m(\mathbf{x})}{C_m(\mathbf{x})} - \mathrm{IRDR}_m \right)^2 \to \min$$

  - **x** is a vector that consists of the weights of words
  - E_m(**x**) is a function that determines the sum of the weights of mis-recognized words
  - C_m(**x**) is a function that determines the sum of the weights of the correct transcript
  - The steepest decent method is adopted to determine the weights that give minimal *F*(**x**)
  - Initially, all weights are set to 1, and then each word weight (x_k) is iteratively updated until the mean square error between WWER and IRDR converges

Experimental Results



Results for whole test-set queries

minimization target in MBR (# of queries)	ASR error rate (%) 1-best → MBR	IRDR (%) 1-best → MBR
WER (287)	21.25 → 20.87	42.67 → 42.65
KER (287)	33.02 → 32.23	42.67 → 42.88
WKER$_{\mathrm{sup.}}$ (287)	38.65 → 38.21	42.67 → **42.46**
WKER$_{\mathrm{semi}}$ (287)	46.43 → 45.97	42.67 → **42.55**

Results for queries whose MBR results differ from 1-best results

minimization target in MBR (# of queries)	ASR error rate (%) 1-best → MBR	IRDR (%) 1-best → MBR
WER (55)	27.24 → 24.86	50.82 → 50.72
KER (50)	40.82 → 35.58	48.13 → 49.59
WKER$_{\mathrm{sup.}}$ (68)	47.96 → 45.43	53.12 → **52.06**
WKER$_{\mathrm{semi}}$ (71)	48.69 → 46.55	48.40 → **47.82**

- Each MBR decoding improved its minimization target
  - Although WER and KER improvement were achieved by MBR, but did not obtain an improvement of IR accuracy
  - On the other hands, according to the minimization of WKER$_{\mathrm{sup}}$. and WKER$_{\mathrm{semi}}$, which are defined with estimated word weights, can achieved an IR performance imporvement

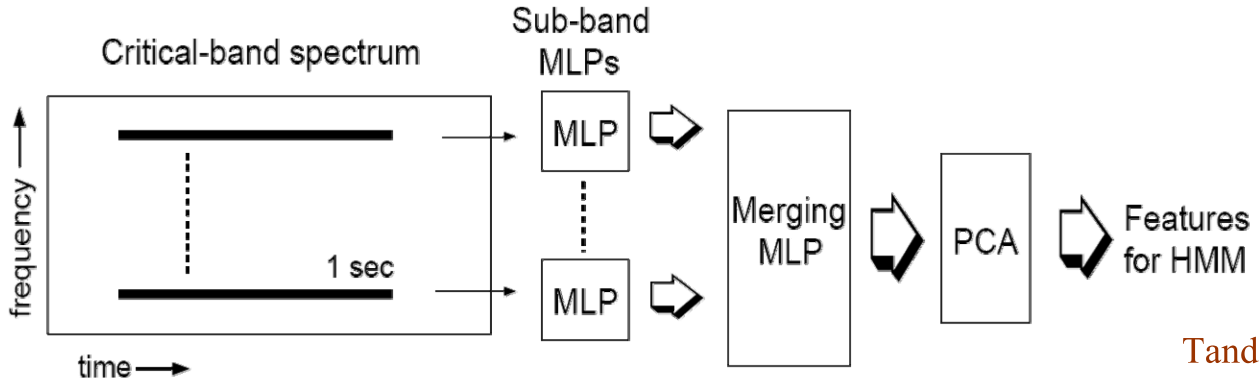
CRANDEM SYSTEMS: CONDITIONAL RANDOM FIELD ACOUSTIC MODELS FOR HIDDEN MARKOV MODELS

Eric Fosler-Lussier, Jeremy Morris

Department of Computer Science and Engineering
The Ohio State University, Columbus, OH
{fosler, morrijer}@cse.ohio-state.edu

Introduction

- In recent years, Conditional Random Fields (CRFs) have been examined as a statistical model for speech recognition
  - Unfortunately, to this point, CRF systems have been used exclusively in the realm of phone classification or phone recognition
    - requires estimation of O(N^2) parameters, where N is the number of state labels
  - In this paper, they explore the use of features derived via CRFs as inputs to a Tandem style HMM ASR system



Tandem System

Deriving Local Posterior Functions for HMMs

- In the Tandem approach, the acoustic input ***X*** is transformed into a more discriminative representation of the input signal via a transformation function ***X'*** = F(***X***) before submitting these features to an HMM system

$$
\begin{aligned}
F(X) &= \mathrm{KLT}(\log P(q_i|X_{i-c}^{i+c})) \\
F(X) &= \mathrm{KLT}(\mathrm{linearize}(P(q_i|X_{i-c}^{i+c})))
\end{aligned}
$$

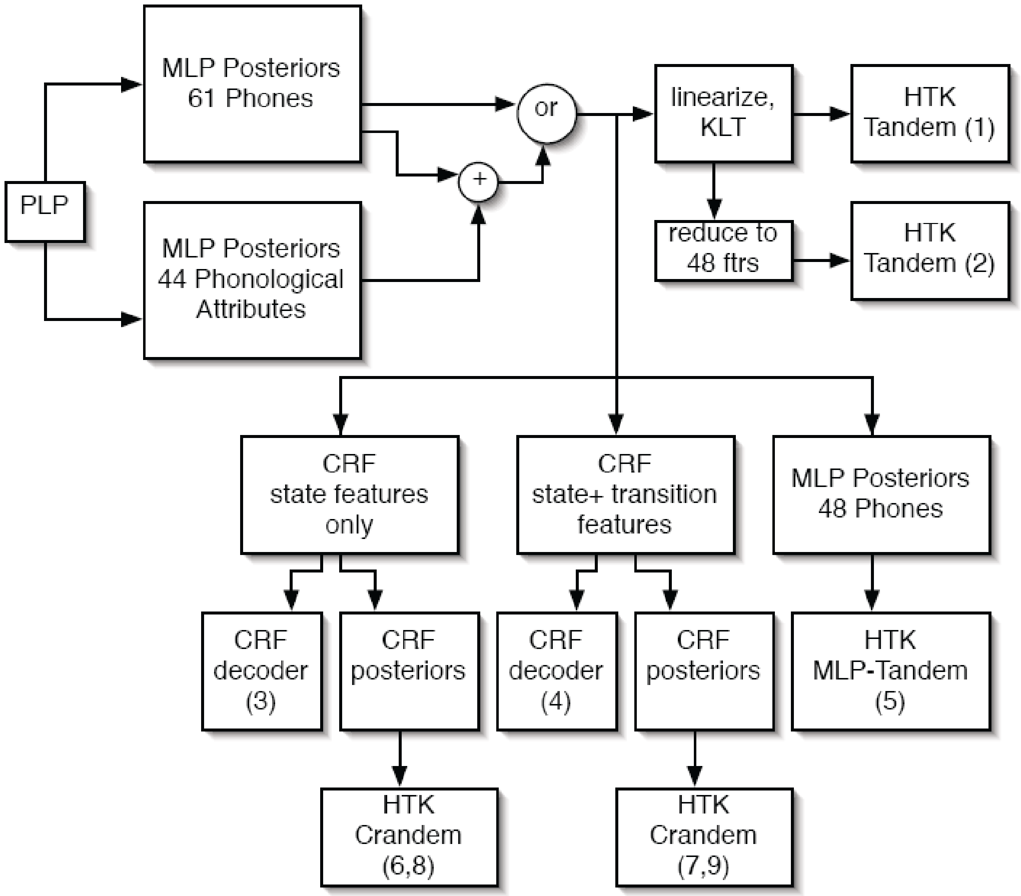
KLT: Karhunen-Loeve transform

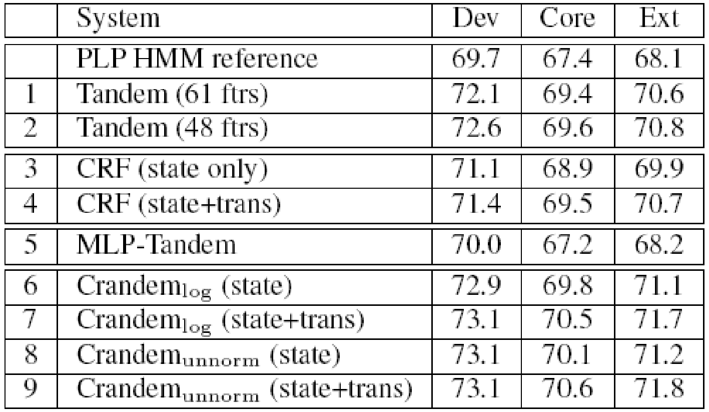
- The transformation *F*(***X***) can be also used in the CRF training paradigm
  - parameters are estimated to maximize the conditional log likelihood of the joint sequence of labels ***Q*** given some representation of the input ***X***

$$P(Q|X) = \frac{\exp(\sum_t \sum_j \lambda_j s_j(q_t, X, t) + \sum_k \mu_k t_k(q_{t-1}, q_t, X, t))}{Z(X)}$$

    - use MLP posterior estimates directly as state feature functions
    - use the self-same MLP posteriors as transition functions

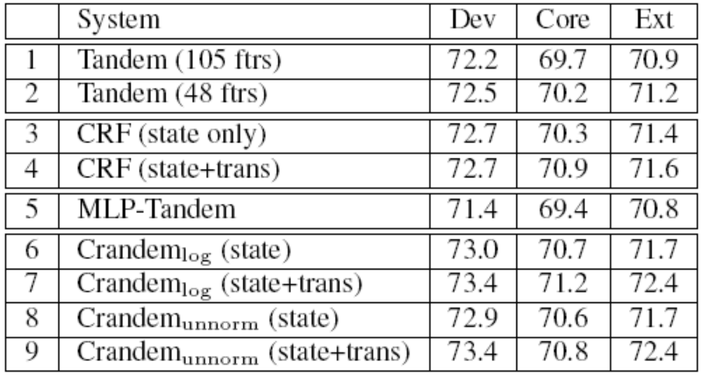
CRANDEM System





	System	Dev	Core	Ext
	PLP HMM reference	69.7	67.4	68.1
1	Tandem (61 ftrs)	72.1	69.4	70.6
2	Tandem (48 ftrs)	72.6	69.6	70.8
3	CRF (state only)	71.1	68.9	69.9
4	CRF (state+trans)	71.4	69.5	70.7
5	MLP-Tandem	70.0	67.2	68.2
6	$\text{Crandem}_{\text{log}}$ (state)	72.9	69.8	71.1
7	$\text{Crandem}_{\text{log}}$ (state+trans)	73.1	70.5	71.7
8	$\text{Crandem}_{\text{unnorm}}$ (state)	73.1	70.1	71.2
9	$\text{Crandem}_{\text{unnorm}}$ (state+trans)	73.1	70.6	71.8

a. System results using 61 phone class posteriors as input



	System	Dev	Core	Ext
1	Tandem (105 ftrs)	72.2	69.7	70.9
2	Tandem (48 ftrs)	72.5	70.2	71.2
3	CRF (state only)	72.7	70.3	71.4
4	CRF (state+trans)	72.7	70.9	71.6
5	MLP-Tandem	71.4	69.4	70.8
6	$\text{Crandem}_{\text{log}}$ (state)	73.0	70.7	71.7
7	$\text{Crandem}_{\text{log}}$ (state+trans)	73.4	71.2	72.4
8	$\text{Crandem}_{\text{unnorm}}$ (state)	72.9	70.6	71.7
9	$\text{Crandem}_{\text{unnorm}}$ (state+trans)	73.4	70.8	72.4

b. System results combining 61 phone class posteriors with 44 phonological feature posteriors

DISCRIMINATIVE FEATURE WEIGHTING USING MCE TRAINING FOR TOPIC IDENTIFICATION OF SPOKEN AUDIO RECORDINGS

Timothy J. Hazen and Anna Margolis

MIT Lincoln Laboratory
Lexington, Massachusetts, USA

Introduction

- Topic identification problem is consisting of two primary stages
  - *feature selection*
    - reduce the large space of potential features to a smaller set which possesses the most relevant or discriminative features for topic ID
      - the mutual information between features and topics, the maximum a posteriori probability of topics given features, or χ^2 statistics
  - *Classification*
    - The use of naive Bayes classifiers is popular throughout much of the topic ID research
      - Because these classifiers use generative models
        - » their training can be performed efficiently
        - » their parameters can be learned and adapted in an on-line fashion
        - » their accuracy is often sufficient for many tasks
      - There are two obvious potential drawbacks to the standard naive Bayes approach
        - » their parameters are generally estimated statistically instead of being trained in a discriminative fashion
        - » the processes of feature selection and model training are generally performed independently instead of jointly
- In this work, we attempt to address the shortcomings of the traditional naive Bayes classifier by applying a discriminative procedure commonly called minimum classification error (MCE) training to the topic ID problem.

Experimental Task Description

- Corpus
  - English Phase 1 portion of the Fisher Corpus
    - 5851 recorded telephone conversations
      - two people were connected over the telephone network and given instructions to discuss a specific topic for 10 minutes
      - Data was collected from a set of 40 different topics
  - In this paper, the corpus was subdivided into four subsets
    - Recognizer training set (3104 calls; 553 hours)
    - Topic ID training set (1375 calls 244 hours)
    - Topic ID development test set (686 calls; 112 hrs)
    - Topic ID evaluation test set (686 calls; 114 hrs)
- Speech Recognizer
  - explore the use of both word-based and phone-based speech recognition
    - each lattice we can compute the posterior probability of any hypothesized word
    - *and expected count* for each word can be computed by summing the posterior scores over all instances of that word over all lattices

Probabilistic Topic Identification

- The goal of topic ID is to determine the likelihood of a document being of topic ***t*** (from a set of topics **T**) given the document's string of words ***W***
- The Naive Bayes Formulation
  - For closed-set topic ID, an audio document will be determined to belong to topic $\boldsymbol{t_i}$ if the following expression holds

$$\forall j \neq i \quad \frac{P(W|t_i)}{P(W|\bar{t_i})} > \frac{P(W|t_j)}{P(W|\bar{t_j})}$$

  - In the naive Bayes approach to the problem, statistical independence is assumed between each of the individual words in ***W***

$$P(W|t) \approx \prod_{i=1}^{N} P(w_i|t) \quad \text{or} \quad P(W|t) \approx \prod_{\forall w \in V} P(w|t)^{C_{w|W}}$$

$$P(W|\bar{t}) = \frac{1}{N_T - 1} \sum_{\forall t_i \neq t} P(W|t_i)$$

  - In practice the score for topic t given words W, expressed as *F(t|W)*

$$\mathcal{F}(t|W) = \sum_{\forall w \in V} C_{w|W} \log \frac{P(w|t)}{P(w|\bar{t})}$$

Probabilistic Topic Identification

- Parameter Estimation
  - The likelihood function *P(w|t)* is estimated from training materials using maximum *a posteriori* probability (MAP) estimation with Laplace smoothing

$$P(w|t) = \frac{N_{w|t} + N_V P(w)}{N_{W|t} + N_V}$$

N_V is the total number of words in the vocabulary
$N_{w|t}$ is the number of times word w occurs in training documents of topic t
$N_{W|t}$ is the total number of words in the training documents of topic t
P(w) represents the prior likelihood of word w occurring independent of the topic

$$P(w) = \frac{N_w + 1}{N_W + N_V}$$

- Feature Selection
  - Select the top *N* words per topic which maximize the posterior probability of the topic ➔ *P(t|w)*

$$P(t|w) = \frac{N_{w|t} + 1}{N_w + N_T}$$

MCE-Based Feature Weighting

- Feature selection can be viewed as a specific case of feature weighting, where each feature receives either a weight of one or a weight of zero
  - In the more general case, we can allow the weights of each feature to be of any value (or at least any positive value)
  - The basic naive Bayes expression can now be generalized to include variable valued features weights

$$\mathcal{F}(t|W) = \sum_{\forall w \in V} \boxed{\lambda_w} C_{w|W} \, f(t|w) \quad \text{where} \quad f(t|w) = \log \frac{P(w|t)}{P(w|\bar{t})}$$

  - The goal is to learn values for the collection of feature weights which minimize the topic ID error rate
    - Use MCE framework to learn the weight

$$\mathcal{M}(W) = \underset{\text{top1}}{\underline{\mathcal{F}(t_I|W)}} - \underset{\text{correct}}{\underline{\mathcal{F}(t_C|W)}}$$

misclassification measure

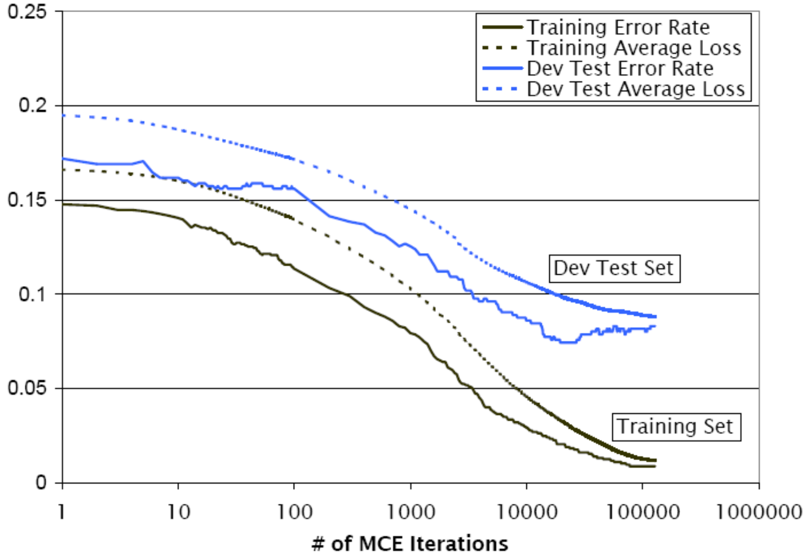
$$\ell(W) = \frac{1}{1 + e^{-\beta \mathcal{M}(W)}}$$

loss function

$$\frac{\partial \ell(W)}{\partial \lambda_w} = \beta \ell(W) \left(1 - \ell(W)\right) \left(f(t_I|w) - f(t_C|w)\right) C_{w|W}$$

gradient

Experimental Results



Experimental Conditions		Topic Error Rate(%)	
Recognition Type	Features	Pre-MCE	Post-MCE
English Words	30373 Words	16.9	7.4
English Words	3155 Words	9.6	7.9
English Phones	13899 3-grams	30.0	19.2
English Phones	3363 3-grams	22.2	21.0
Hungarian Phones	14413 3-grams	65.0	48.5
Hungarian Phones	3494 3-grams	53.0	47.7

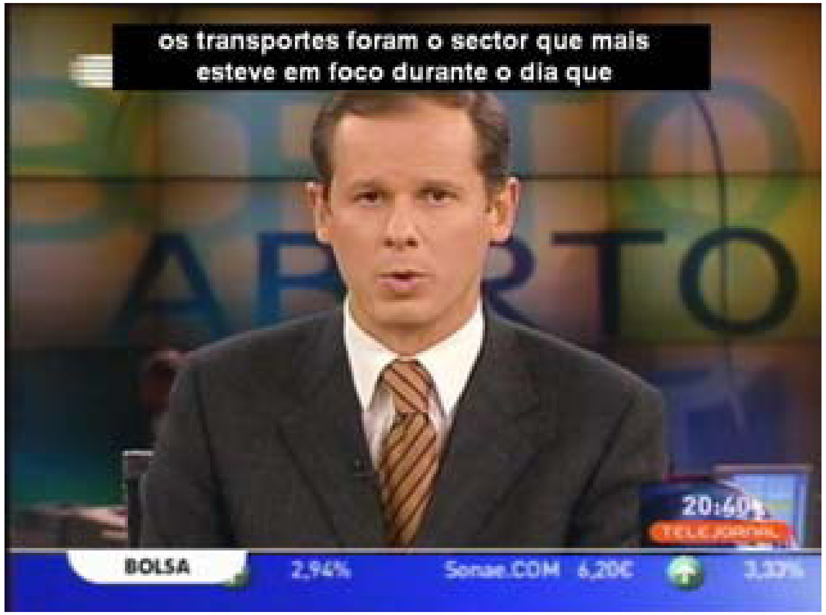
BROADCAST NEWS SUBTITLING SYSTEM IN PORTUGUESE

J. Neto(1,2), *H. Meinedo*(1), *M. Viveiros*(1), *R. Cassaca*(1), *C. Martins*(1), *D. Caseiro*(1,2)

(1) L2F – Spoken Language Systems Lab / INESC-ID
(2) Instituto Superior Técnico / Technical University of Lisbon
Joao.Neto@inesc-id.pt
http://www.l2f.inesc-id.pt

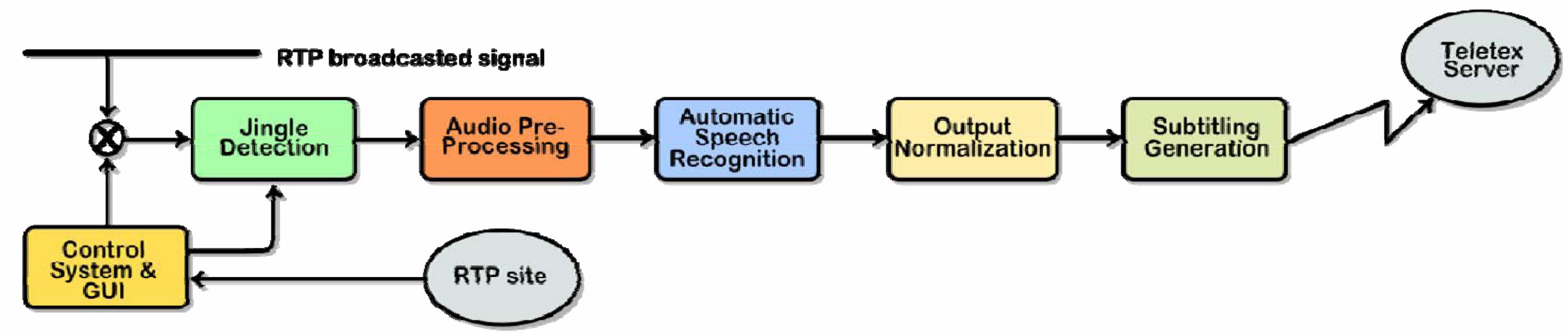
Introduction

- The subtitling of broadcast news (BN) programs are starting to become a very interesting application
  - due to the technological advances in Automatic Speech Recognition (ASR) and associated technologies as Audio Pre-Processing (APP)
- Who or what can get benefit from subtitling
  - hearing handicapped, elderly people, people in noisy places, content search, selective dissemination of information and machine translation

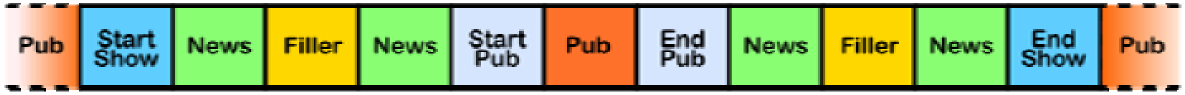


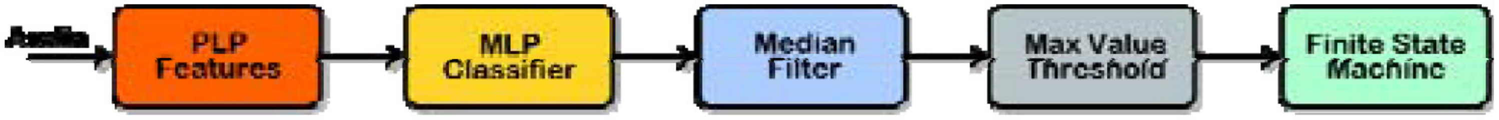
os transportes foram o sector que mais
esteve em foco durante o dia que

Block Diagram of the Subtitling System



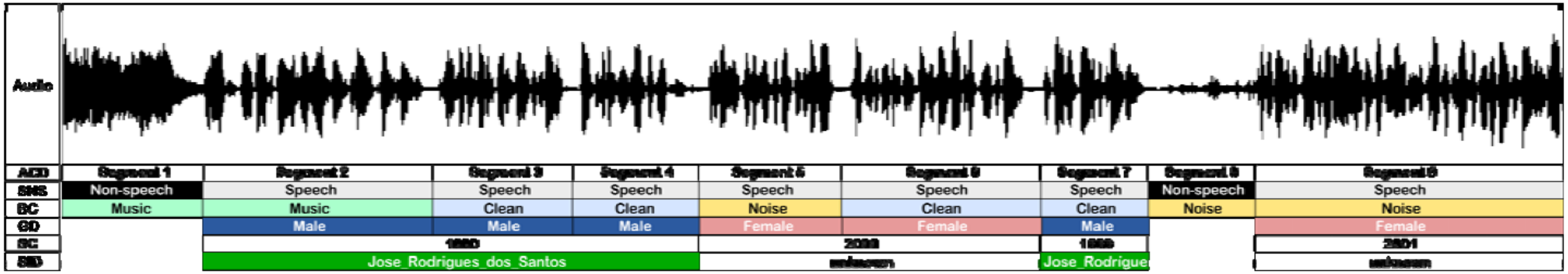
- Jingle Detection
  - "Jingles" and are used in Broadcast News shows for drawing the listener's attention to important events like the start and the end of the show
    - The goal of this block is to identify, in the audio stream, specific acoustic patterns
    - The Jingle Detection block also filters the commercials and the end jingle





Block Diagram of the Subtitling System (cont.)

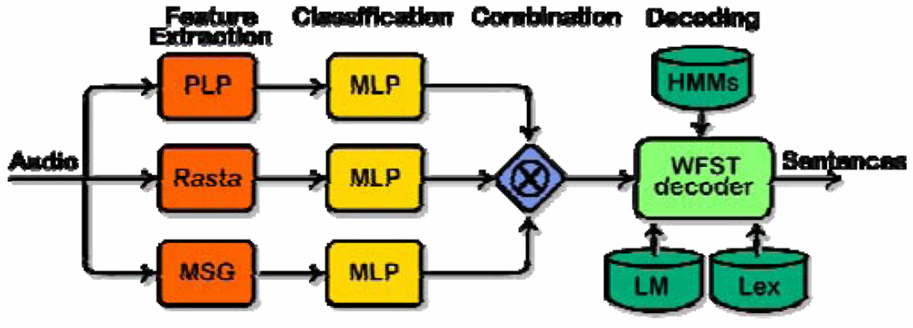
- Audio Pre-Processing (APP)
  - The operation of the APP block is two-fold
    - to filter the non-speech parts
    - to give additional information to the following blocks
      - Gender classification, Background classification, Speaker clustering, Speaker Identification
  - This block contains three classifier
    - Audio segmentation, Audio classification, Speaker classification



ACD	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6	Segment 7	Segment 8	Segment 9
SNS	Non-speech	Speech	Speech	Speech	Speech	Speech	Speech	Non-speech	Speech
BC	Music	Music	Clean	Clean	Noise	Clean	Clean	Noise	Noise
GD		Male	Male	Male	Female	Female	Male		Female
SC		1000			2000		1000		2001
SID		Jose_Rodrigues_dos_Santos			unknown		Jose_Rodrigue		unknown

Block Diagram of the Subtitling System (cont.)

- Automatic Speech Recognition (ASR)
  - based on a hybrid speech recognition structure combining the temporal modeling capabilities of Hidden Markov models (HMM), with the pattern discriminative classification capabilities of MLPs



- Output Normalization and Subtitling Generation
  - improve the readability of the subtitles