What's Happening In Accents & Dialects ?

A Review Of The State Of The Art (post-Interspeech 2013)

Martin Russell – University of Birmingham
Andrea DeMarco – University of East Anglia
Christophe Veaux – University of Edinburgh
Maryam Najafian – University of Birmingham

UKSpeech

UNIVERSITY OF BIRMINGHAM University of East Anglia THE UNIVERSITY of EDINBURGH

Overview of Themes

1) Classification & Identification – Andrea
2) Speech Synthesis - Christophe
3) Automatic Speech Recognition – Maryam
4) Human Perception and Production – Maryam

Classification & Identification

- Languages, accents & dialects
- A total of 11 papers surveyed (not a lot)
- Various application scenarios, but most work is on Language Identification (LID)
- We'll have a look at:
  - Feature extraction techniques
  - Classification methods
  - Corpora
  - Results
  - What's happening next

Classification & Identification - Application Scenarios

- Foreign Accent Detection from Spoken Finnish [5]
- Native British Accent Classification [7]
- Accent Quantification of Indian Speakers of English [11]
- Language Identification [1,2,3,4,6,8,9,10]

Classification & Identification - Feature Extraction

- MFCC → RASTA → CMVN → VTLN → SDC
- MFCC → Warping $X \sim N(0,I)$ → SDC → Concatenate
- MFCC → Delta → Delta-Delta → CMVN
- Phone lattices and n-grams, absolute (what) and relative (where) distance kernels (PARF)
- Phone Log-Likelihood Ratios (PLLR) → PCA
- Phonotactic i-Vectors

Classification & Identification - Classification Methods

- i-Vectors – a point estimate of an utterance in variability subspace
- Speaker Compensation
  - Linear/Semi-supervised/Heteroscedastic/Probabilistic Discriminant Analysis
  - Neighbourhood Component Analysis
- Binary Genetic Algorithm-based classifier fusions
- Traditional GMM models for supervised phoneme classes
- SVM Kernels
- DARPA RATS ANN on i-vectors
  - 3 layers, i-vector input, 6-language posterior output
  - 400-700 hidden nodes
- DARPA RATS Adaptive Gaussian Backend

Classification & Identification - Corpora

- FSD (Finnish National Foreign Language Certificate Corpus)
- ABI (Accents of the British Isles Corpus)
- Custom Indian Speaker Dataset
- NIST Language Recognition Evaluation (LRE)
- RATS LID Data Corpus (5 targets, 10 non-targets)

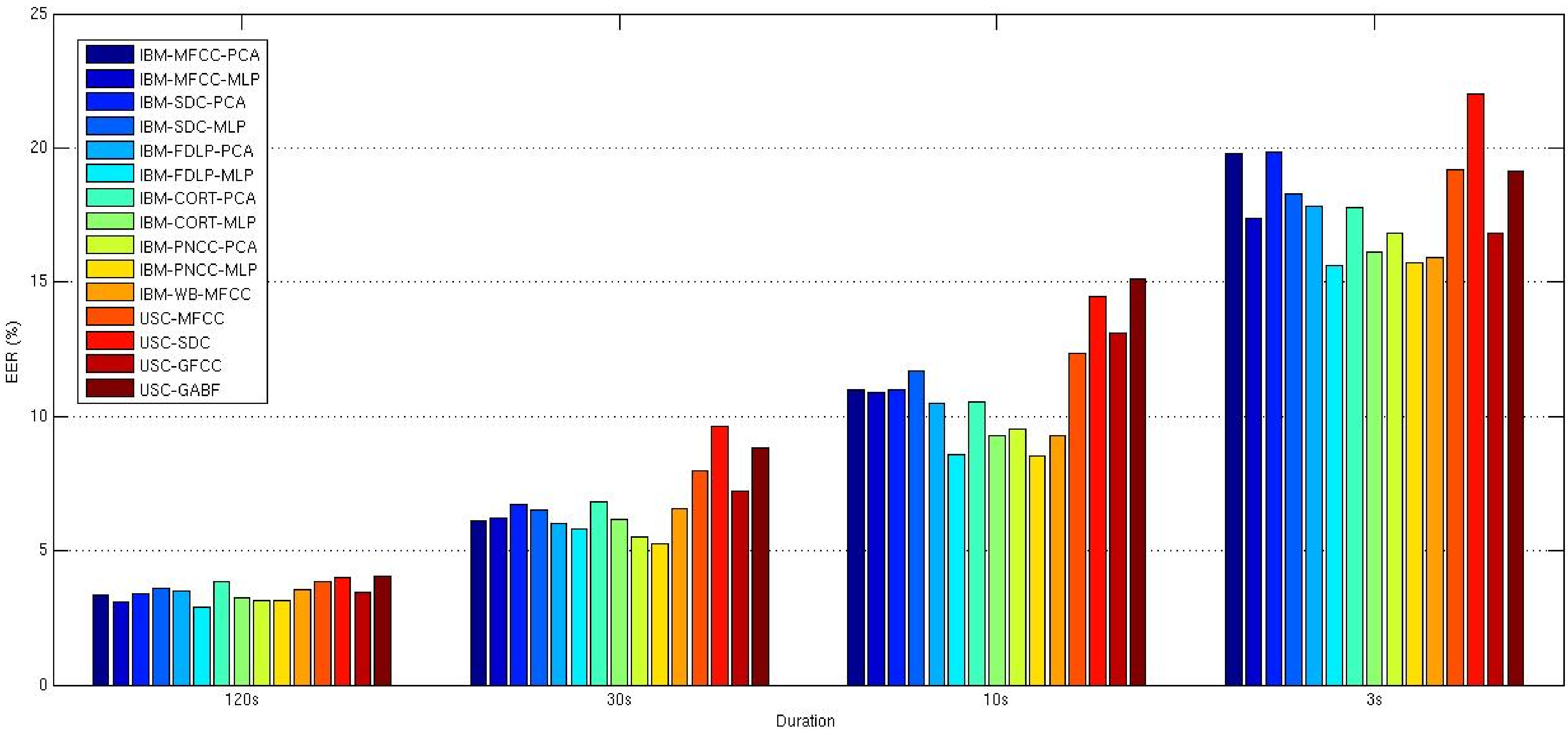
Classification & Identification - Results

Corpus	Novel Method	Baseline
FSD (iVector)	20.01% EER	24.13% EER
ABI (iVector)	81% Accuracy	73.6% Accuracy
LRE (PARF)	19.89% EER (3s test)	23.90% EER (3s test)
LRE (PLLR)	3.21% C_{avg} ,1.79% C_{avg}	3.79% C_{avg} ,2.09% C_{avg}
RATS (iVector-ANN)	6.95% EER	8.99% EER
LRE (Phon. iVector)	19.11% EER (3s test)	22.60% EER (3s test)
RATS (iVector-AGB)	3.6% C_{avg} (30s test)	4.9% C_{avg} (30s test)

- Indian accent strength (like in other languages) can be tied down to models of specific phonemes – mostly consonants in Indian. Machine performance equalled human listeners.

Take Home Message (1)

- Feature Vector Overview for TRAP Language Identification System for RATS Phase II Evaluation



Take Home Message (2)

- Different factor sizes/UBM components/Dim Reductions. Classifiers behave differently – Fusion gives a big boost (Accents of the British Isles)

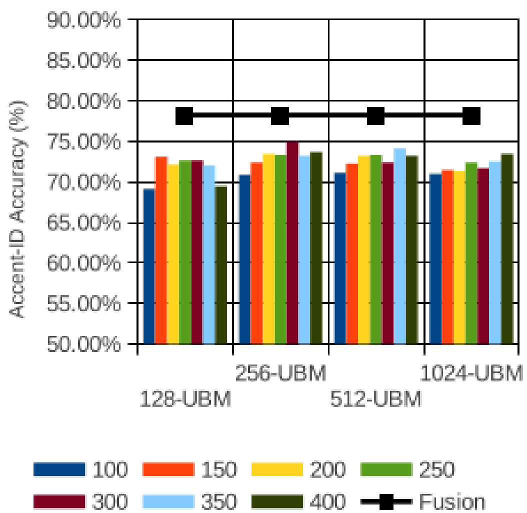


Figure: LDA Classification over R-LDA reduced I-vectors.

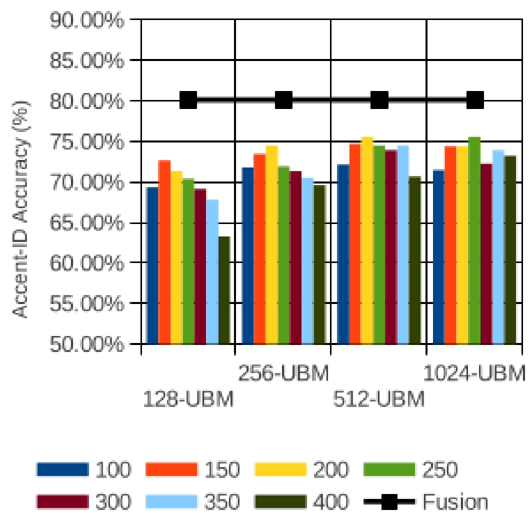


Figure: LDA Classification over SDA reduced I-vectors.

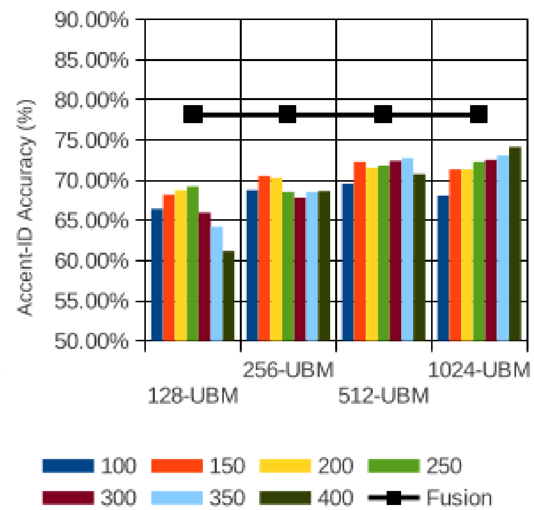


Figure: SVM Classification over SDA reduced I-vectors.

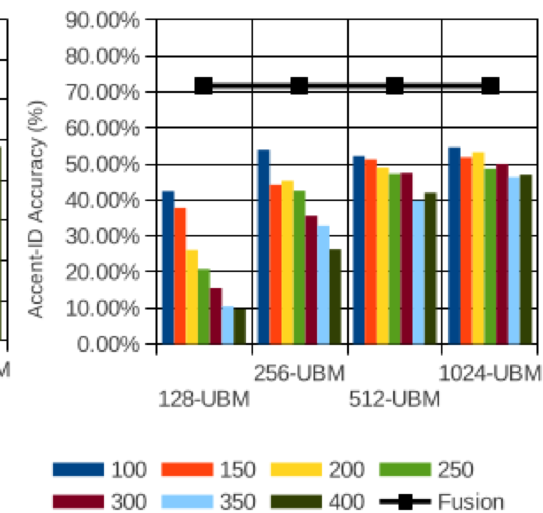
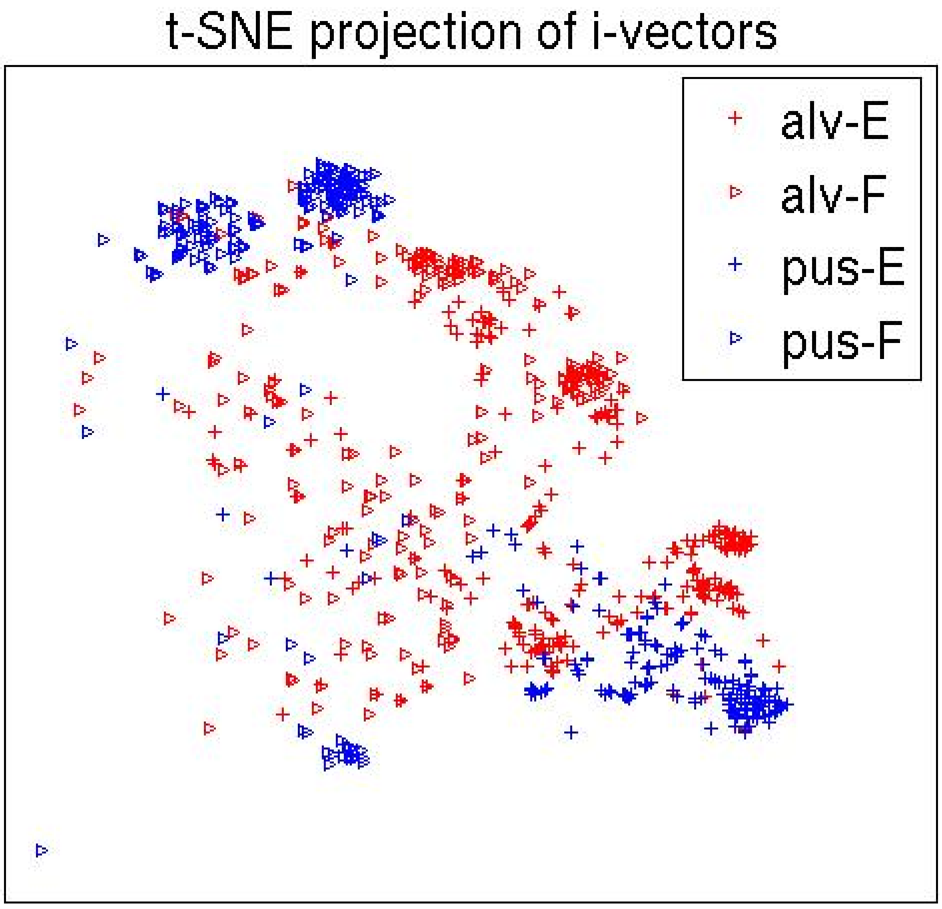
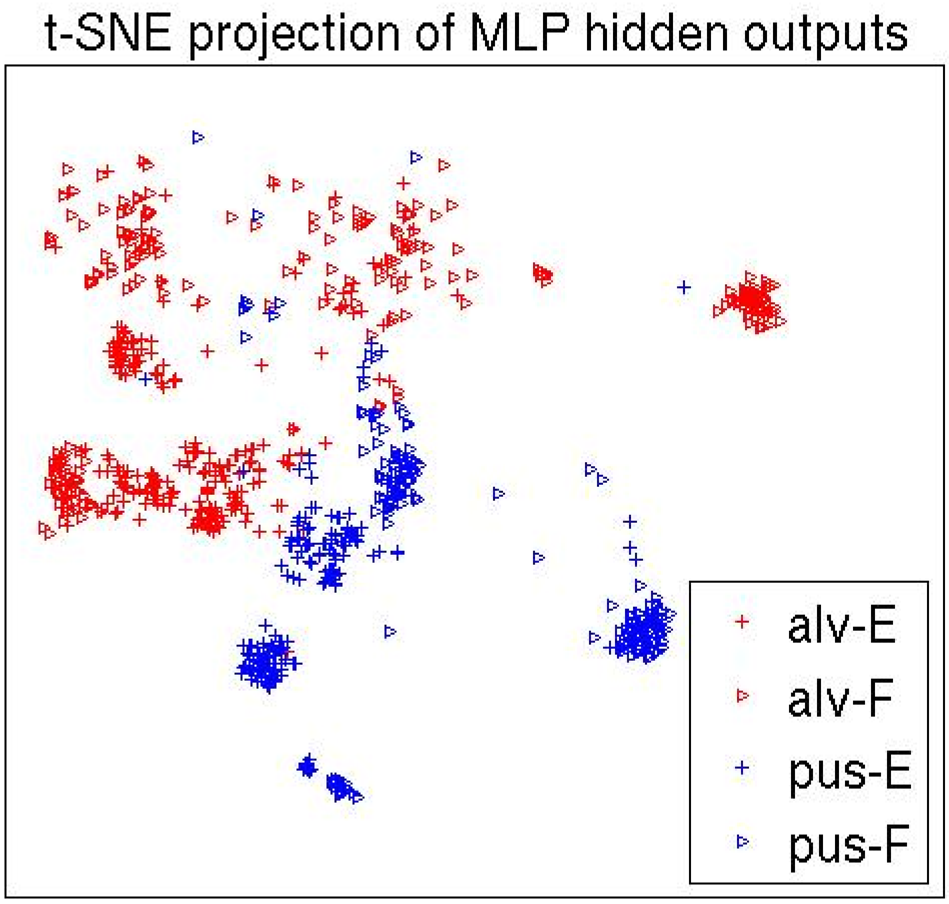


Figure: 1-NN Classification over NCA reduced I-vectors.

Take Home Message (3)

- Stochastic Neighbour Embedding (SNE) Mapping of I-vectors (Language Identification System for RATS Phase II Evaluation)





Conclusions

- Work is still traditionally split between acoustic-only and acoustic-phonetic classification.
- Most of the work is in acoustic-only methods.
- Interspeech 2013 – Capitalize on I-vectors
- Interspeech 2014 – A move towards Artificial Neural Networks/Deep Belief Networks instead of/added on to current scoring methods?

References

1. ***'Parallel Absolute-Relative Feature Based Phonotactic Language Recognition'***, Weiwei Liu et. al.

2. ***'Dimensionality Reduction of Phone Log-Likelihood Ratio Features for Spoken Language Recognition'***, Mireia Diez et. al.

3. ***'Improvements in Language Identification on the RATS Noisy Speech Corpus'***, Jeff Ma et. al.

4. ***'Regularized Subspace n-Gram Model for Phonotactic iVector Extraction'***, Mehdi Soufifar et. al.

5. ***'Foreign Accent Detection from Spoken Finnish Using i-Vectors'***, Hamid Behravan et. al.

6. ***'Adaptive Gaussian Backend for Robust Language Identification'***, Mitchell McLaren et. al.

7. ***'Native Accent Classification via I-Vectors and Speaker Compensation Fusion'***, Andrea DeMarco et. al.

8. ***'The Albayzin 2012 Language Recognition Evaluation'***, Luis Javier Rodriguez-Fuentes et. al.

9. ***'TRAP Language Identification System for RATS Phase II Evalutation'***, Kyu J. Han et. al.

10. ***'Improving Language Identification Robustness for Highly Channel-Degraded Speech Through Multiple System Fusion'***, Aaron Lawson et. al.

11. ***'Automatic Accent Quantification of Indian Speakers of English'***, Jian Cheng et. al.

Accents & dialects in TTS / Selected topics

- TTS in various accents / dialects
  - ⇨ *Personalisation of speech synthesis (encourages interaction)*
- Accent conversion / interpolation
  - ⇨ *Computer aided language learning (self reference)*

Belongs to more general topics:

- TTS for under-resourced languages
- Cross-lingual speaker adaptation for TTS

Accents / Challenges for TTS

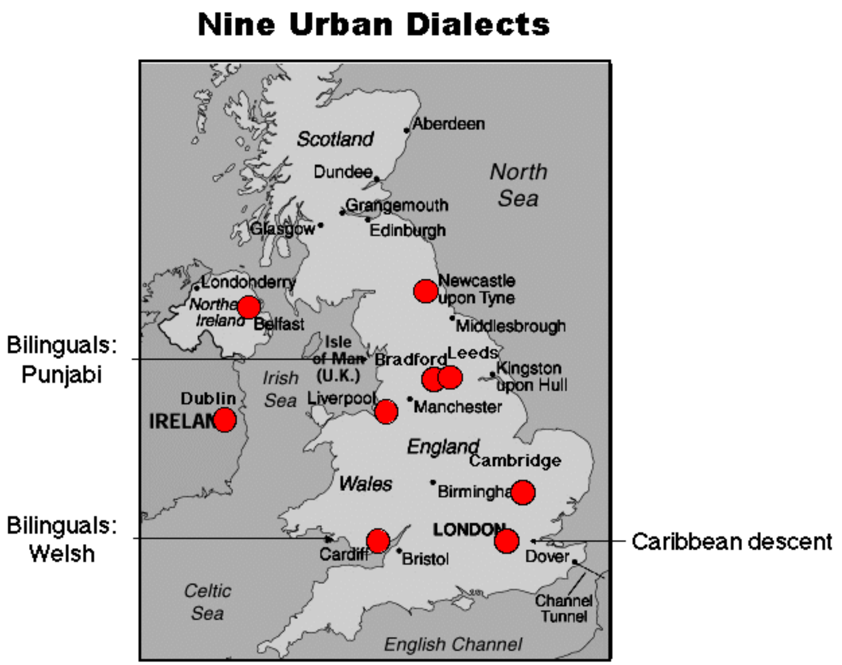
- Accent types
  - Geographical, Sociological, Foreign accent
  - ⇨ *may be difficult to define (discrete vs continuum or mixed)*
- Accent variation
  - Not just a shift in phonetic realisation
  - Change of phonetic inventory
  - Phonological variation can spread over segments
  - Change of segmental structure (insertion/deletion)
  - Intonational variation
  - ⇨ *adaptation of the phone models is not enough*

Dialects / Challenges for TTS

- Types
  - Geographical
  - Sociological (sociolects)

  Normally seen as discrete

  but may be continuous [Saussure]



- Dialect variation
  - change in lexical and grammatical structure (+ accent variation)

  Linguistic knowledge required

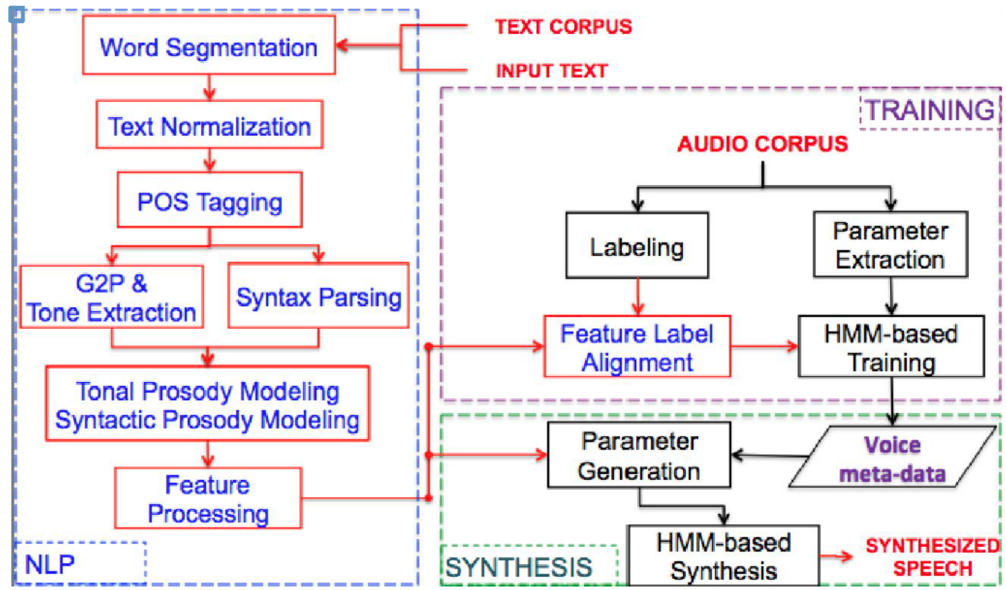
  ➡ same situation as under-resourced languages

Scenarios

- TTS in various accents & dialects
  - Fully resourced accent/dialect
  - Under-resourced accent/dialect/language
- Accent / Dialect conversion or interpolation
  - Accent conversion
  - Accent interpolation
  - Cross-lingual speaker adaptation

Fully resourced accent/dialect

- HMM-based TTS for Hanoi Vietnamese [Nguyen, 2013]
  - NLP module
    - Phonetic inventory
    - Phonological features
    - Lexicon
    - G2P and POS Tagger

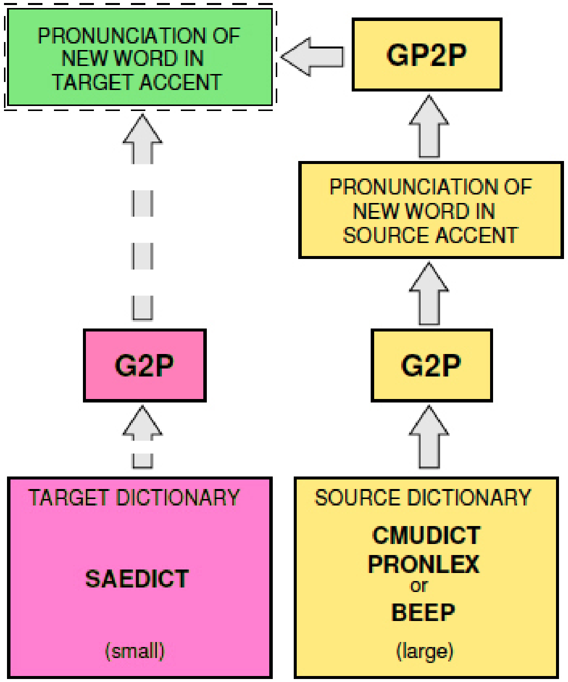


  - Training of HMM-based synthesizer on a dialectal corpus
    - VNSpeechCorpus (Hanoi Vietnamese, 630 sentences)
- Advanced Lexicons (Unilex, Combilex) [Richmond, 10]
  - Encode different pronunciations based on morphological derivation

Under-resourced accent/dialect/language

- Learning G2P requires a large training set

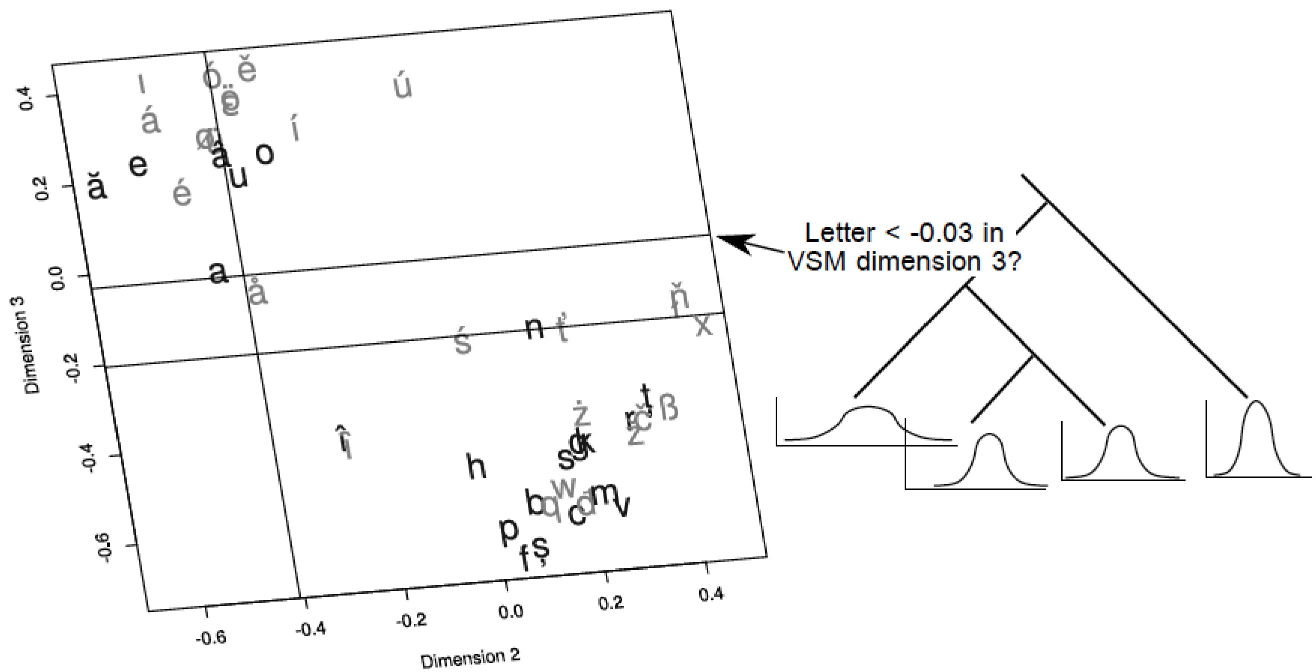
➡ Decision-tree based conversion of pronunciations dictionnary from one accent to another [Loots, 10]



➡ Iterative refinement of G2P system using a small lexicon as bootstrap [Goel, 10] (ASR)

Under-resourced accent/dialect/language

- Build a TTS system with little or no supervision [Watts,13]
  - Unsupervised linguistic representation learned from text
    - Vector Space Model used to characterise ‘textual units’



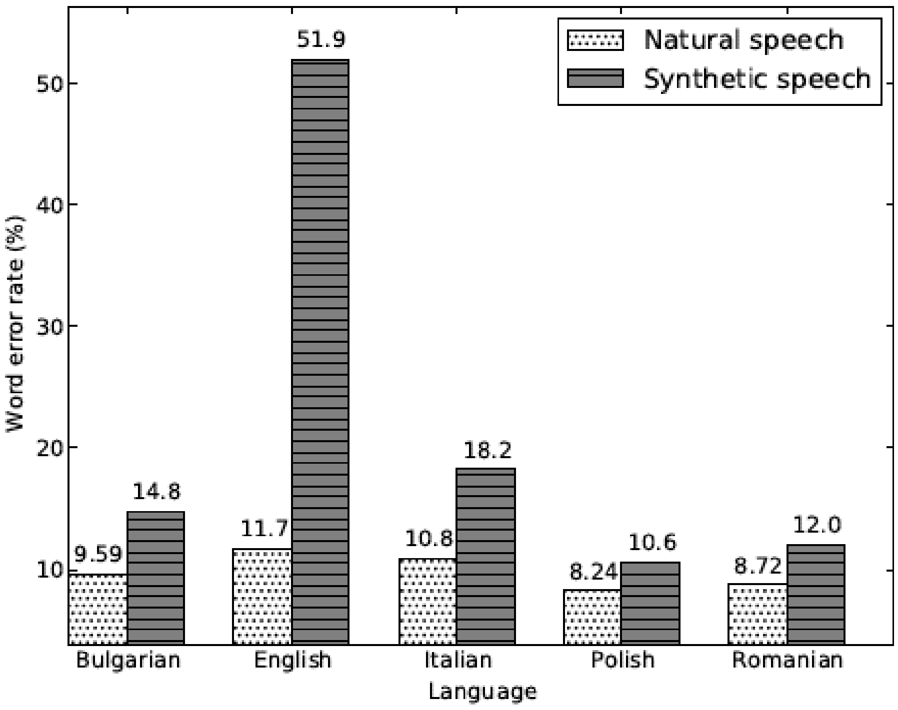
  - Letter based speech modeling units instead of phonemes

Under-resourced accent/dialect/language

- Build a TTS system with little or no supervision
  - Lightly supervised alignement [Stan, 13]
    - Graphem models instead of phone models
    - Discriminative training (Maximum Mutual Information)

- Corpus of "found speech"
  - Audiobooks, 14 languages

➡ Performs well if relatively simple relation between graphemes and phonemes

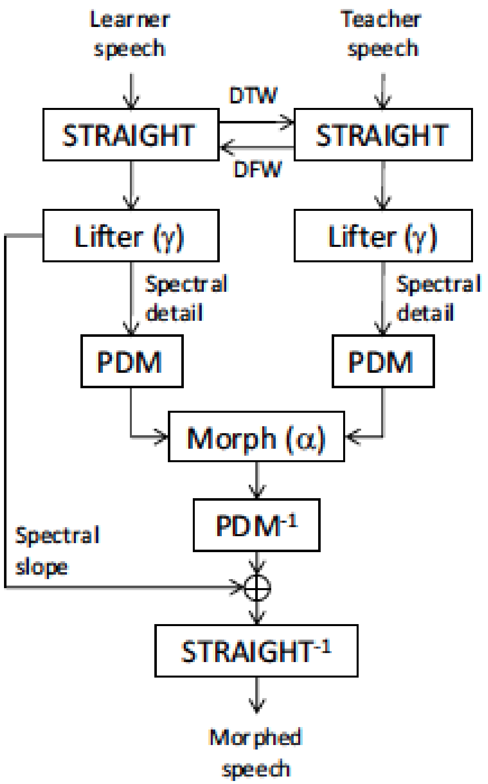


Scenarios

- TTS in various accents & dialects
  - Fully resourced accent/dialect
  - Under-resourced accent/dialect
- Accent / Dialect conversion or interpolation
  - Accent conversion
  - Accent interpolation
  - Cross-lingual speaker adaptation

Accent conversion

- Voice Morphing strategy
  - Foreign accent removal [Aryal,13]
    - Separation of **spectral slope** and **spectral fine details**
    - Spectral details represented by pulse density modulation (PDM)
    - Interpolation of the PDM representations
  - Formant-based VTLN [Qian,11]



- Speaker adaptation strategy [Karhila,11]
  - **Rapid adaptation** of accent specific average voices models using limited amount of speaker’s data (5 to 15 sentences)

Accent interpolation

- HMM linear interpolation [Astrinaki, 13]
  - Clusters of speakers with same accent
  - Interpolation between these clusters

- Constrained HMM interpolation

Different interpolation modes [Pucher, 10]

•Simple linear interpolation

•Discrete phonological shifts

  - Add a **switching rule** to control the HMM interpolation

•Segmental structure changes (insertion/deletion)

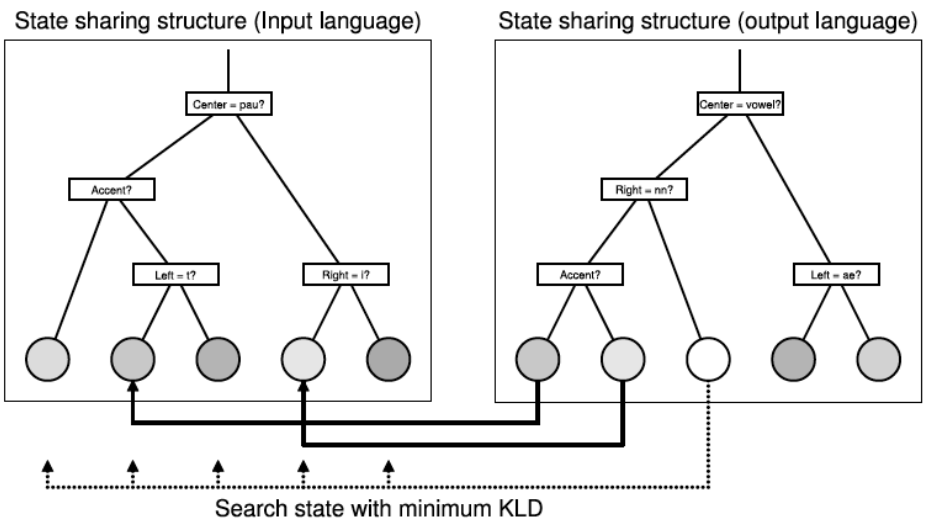
  - use of **null phones** which correspond to a phone with zero duration

One-line interpolation since the choice of the interpolation mode depend on the context.

Cross-lingual speaker adaptation

(Speaker A, L1) ➡ (Speaker A, L2)

- Unsupervised state-level mapping [Oura, 10]



- KLD mapping between “similar” states of average voices models (L1, L2)
- State-dependent transforms are generated using the L1 average voice model and the speaker data
- These transforms are applied to the states of the L2 average voice in order to generate the speaker’s model in L2.

- Structural KLD mapping [Toman, 13]
  - Modified KLD mapping is dependent of phonological context.
  - Used for **cross-dialect adaptation**.

[Saussure] ‘***Course in General Linguistics’,*** Duckworth, London (1983), Original work 1916

[Nguyen, 13] ***'HMM-Based TTS for Hanoi Vietnamese: Issues in Design and Evaluation'***, Nguyen et. al., Interspeech 2013

[Stan, 13] “***TUNDRA: A Multilingual Corpus of Found Data for TTS Research Created with Light Supervision'***, Adriana Stan et. Al, Interspeech 2013

[Loots, 10] ***'Automatic conversion between pronunciations of different English accents’***, Linsen Loots et al, Speech Communication, 2011

[Goel, 10] ***’Approaches to automatic lexicon learning with limited training examples’***, Goel et al, Icassp 2010.

[Watts, 13] ***’Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from ‘found data’: evaluation and analysis'***, Oliver Watts et. Al, SSW8, 2013

[Stan, 13] ***’Lightly supervised Discriminative Training of Grapheme Models for Improved Sentence-level Alignment of Speech and Text Data'***, Adriana Stan et. al., Interspeech 2013

[Aryal, 13] ***‘Foreign Accent Conversion through Voice Morphing'***, Aryal et. al., Interspeech 2013

[Karhila, 11] ***’Rapid Adaptation of Foreign-accented HMM-based Speech Synthesis'***, Karhila et. al., Interspeech 2011

[Qian, 11] ***A frame mapping based HMM approach to cross-lingual voice transformation, Qian et al., Icassp 2011***

[Astrinaki, 13] ***’Reactive accent interpolation through an interactive map application'***, Astrinaki et. al., Interspeech 2013

[Pucher, 10] ***‘Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis'***, Pucher et. al., Speech Communication, 2010

[Oura, 10] ***‘Unsupervised Cross-lingual speaker adaptation for HMM-based Speech Synthesis'***, Oura et. al., Icassp 2010

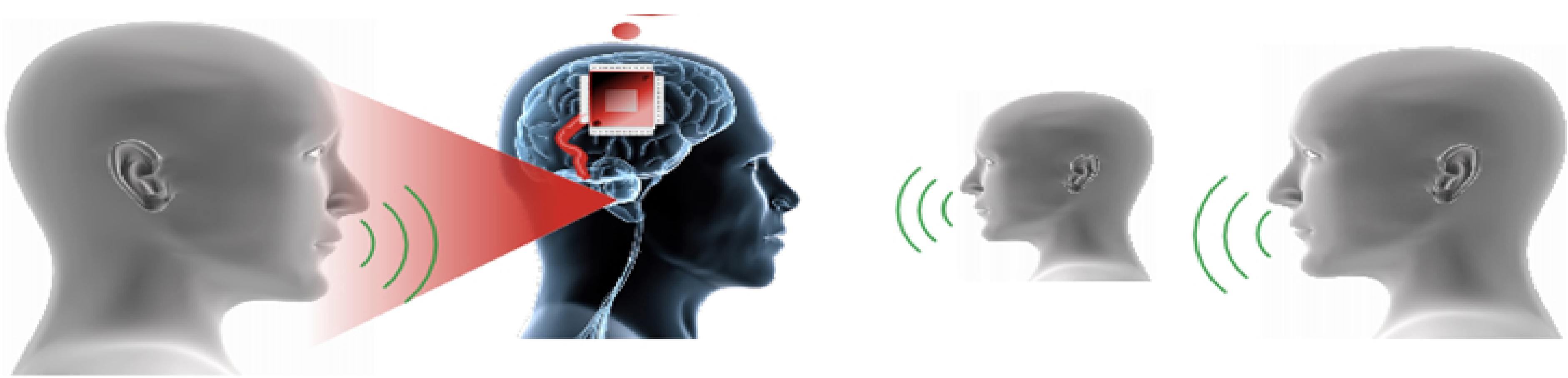
[Toman, 13] ***‘Structural KLD for Cross-Variety Speaker Adaptation in HMM-based Speech Synthesis'***, Toman et. al., IASTED 2013

Overview

- Human Perception and Production
- Automatic Speech Recognition (ASR)

Human Perception & Production

- Cross language effects:
  - Effect of L1(native) speech production in L2 (foreign) perception
  - Effect of L2 speech production in L1 perception
- Effect of accents on human perception
- Multi-lingual/multi-accent cocktail party



Effect of L1 in L2 perception

- Vowels with particular acoustic properties perceived differently according to listener's native language (Italian listener-US English vowels) [10]
- Role of L1 phonology in L2 perception [11], [12]
  - Vowel devoicing in Japanese carried over to German, leading to perceptual difficulties for native German listener
  - Spanish speakers perceiving for two French vowels

Effect of Accent on Human Perception (1)

- Interference between perception of regional accent and speech disorder [6]
  - Disordered speech: weak influence of regional accent on perception of speech disorder
  - Accented speech: listeners unfamiliar with a regional accent may perceive accent differences as a slight speech disorder (when none is present)

Effect of Accent on Human Perception (2)

- An unfamiliar accent slows down spoken word recognition for native and non-native listeners (Australian-, Jamaican-, Cockney-accented English/eye-tracker experiments) [7,8]
  - Category Shifting (CS) differences caused more distraction than Category Goodness (CG)
  - CG: A2 phones constitute 'deviant' from that of A1
  - CS: A2 phones cross A1 phonological boundary

Multi-lingual/-accent cocktail party

- Intelligibility at a multi-accent cocktail party [13]
  - More interference when the target and the masker shared common dialect features
  - More interference when listeners heard their own dialect in the masking babble
- Intelligibility at a multi-lingual cocktail party [14]
  - Acoustic and linguistic information from babble spoken in a known language to the listener competed with the target words
  - Whereas for babble produced in unknown languages only acoustic information was involved

ASR and Accents/Languages

- ASR research focus at Interspeech 2013 is on Deep Belief Networks
- Focus in this talk is on explicit methods to accommodate accent

Spoken Dialect is Mixture of Various Dialects

- Spoken dialects treated as a mixture of various dialects [1]
- Estimation of speaker specific-mixing ratio for Japanese dialects
  - Simple Counting: Count dialect-specific pronunciations to estimate pronunciation dictionary mixture weights
  - Topic-modelling: Categorise words into topics with different dependencies on dialects (Language Model)
  - Topic modelling gives slightly better results

General, accent- and speaker-specific Polyphone Decision Trees(PDTs)

- Recognition of South-Asian accented English[2]
- Comparison of WERs for PDTs trained on general, SoA and speaker-dependent data
- Comparison of distance between PDTs
  - For 'small' PDTs (1k GMMs) SD better that AD better than baseline
  - Little difference in performance for larger PDTs (3k GMMs), despite significant dissimilarity between the trees

Under-resourced / Cross-lingual ASR

- Use of cross-lingual SGMM and Tandem features outperforms conventional HMM/GMM-MFCC for under-resourced languages [3]
- Improve performance on the target language by initializing/training it with a multilingual multilayer perceptrons (MLPs) [4]

Training Data Selection

- How can we get the best performance with the smallest amount of training data (for example, for accented speech)
- iVector-based method for acoustic data selection from a large corpus [5]
- Proposed approach outperforms random data set selection

- Should we continue research in explicit/dialect adaptation?
- Or, will DBNNs solve the problem for us?

Interspeech 2013 References

1. **Automatic Estimation of Dialect Mixing Ratio for Dialect Speech Recognition** *Naoki Hirayama, Koichiro Yoshino, Katsutoshi Itoyamar, Shinsuke Mori, Hiroshi G. Okuno*

2. **Accent- and Speaker-Specific Polyphone Decision Trees for Non-Native Speech Recognition** *Dominic Telaar, Mark C. Fuhs*

3. **Crosslingual Tandem-SGMM: Exploiting Out-Of-Language Data for Acoustic Model and Feature Level Adaptation** *Petr Motlicek*1, *David Imseng, Philip N. Garner*

4. **Multilingual Multilayer Perceptron For Rapid Language Adaptation Between and Across Language Families** *Ngoc Thang Vu and Tanja Schultz*

5. **iVector-based Acoustic Data Selection** *Olivier Siohan, Michiel Bacchiani*

6. **Perceptual interference between regional accent and voice/speech disorders** *Alain Ghio, Médéric Gasquet-Cyrus, Juliette Roquel, Antoine Giovanni*

7. **L2 English Learners' Recognition of Words Spoken in Familiar versus Unfamiliar English Accents** *Jia Ying, Jason A. Shaw, Catherine T. Best*

8. **Recognizing words across regional accents:The role of perceptual assimilation in lexical competition** *Catherine T. Best*, *Jason A. Shaw , Elizabeth Clancy*

9. **The Effects of Perceptual and/or Productive Training on the Perception and Production of English Vowels /ɪ/ and /iː/ by Cantonese ESL learners** *Janice Wing Sze Wong*

10. **Salento Italian listeners' perception of American English vowels** *Bianca Sisinni*1, *Paola Escudero*2 and *Mirko Grimaldi*

11. **On the role of L1 speech production in L2 perception: Evidence from Spanish learners of French** *Natalia Kartushina* 1, *Ulrich Hans Frauenfelder*1

12. **Architekt or Archtekt? Perception of devoiced vowels produced by Japanese speakers of German** *Frank Zimmerer, Rei Yasuda, Henning Reetz*

13. **Regional accents affect speech intelligibility in a multitalker environment** *Ewa Jacewicz, Robert Allen Fox*

14. **Intelligibility at a multilingual cocktail party: Effect of concurrent language knowledge** *Aurore Gautreau, Michel Hoen Fanny Meunier*