Improving Pronunciation Modelling in Automatic Speech Recognition (ASR)

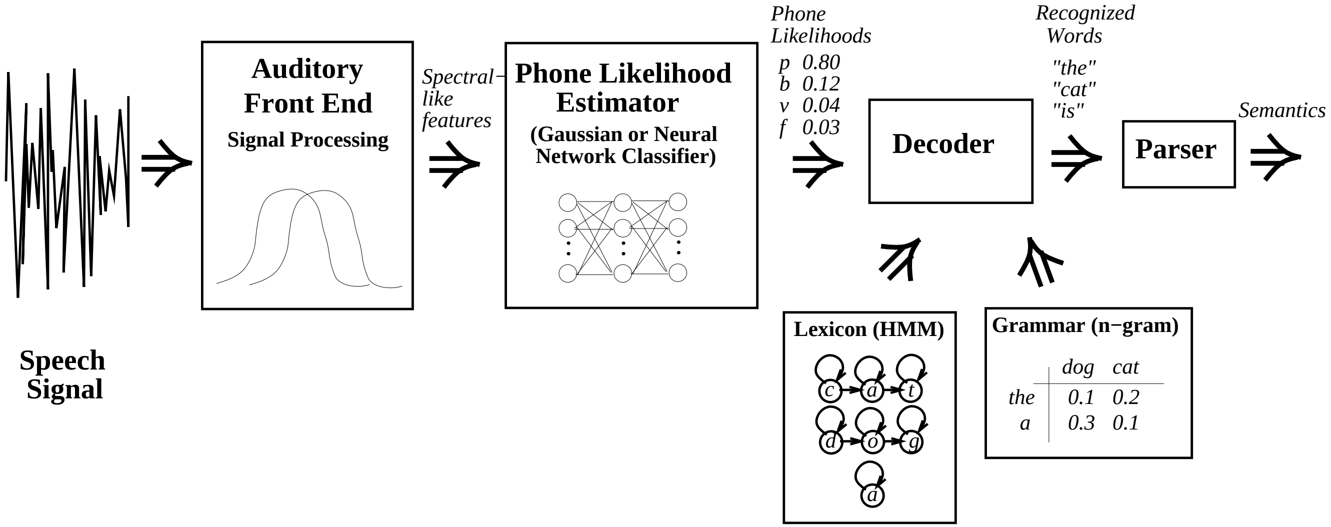
Mathew Magimai Doss and Hervé Bourlard

{mathew, bourlard}@idiap.ch

IDIAP Research Institute, Martigny, Switzerland

Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

Standard ASR Approach



Lexicon should reflect

- Intra- and inter-speaker variability
- Lexical variability (coarticulation, assimilation)

Standard Pronunciation Modeling

Lexical model = first-order Markov model/graph of phonetic units:

- Standard lexical dictionary
- Knowledge-based, e.g. enriched by applying phonological rules
- Data-driven, e.g., MM inference from recognition output followed by HMM retraining
- Mix of the above.

Goal of this work

- Evaluating the “stability” of baseform pronunciations.
- Improving “stability” of pronunciation models by introducing “auxiliary variables”.
- Evaluate lexical models without looking at recognition rates.

Auxiliary Variables

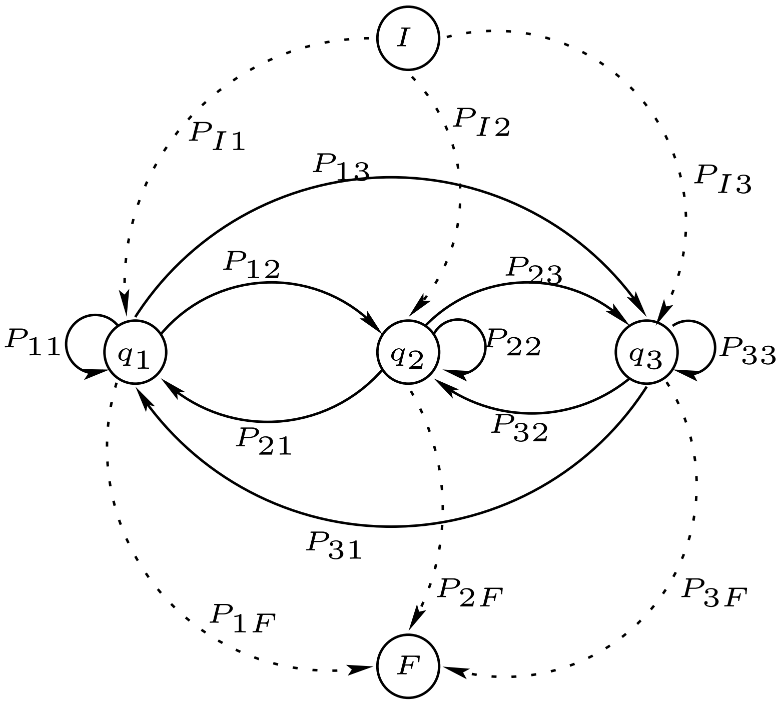
- Instead of:
  - Changing the acoustic features and/or
  - Changing the baseform topologies
- Add a conditional (auxiliary) variable, a, to the HMM emission PDF, i.e.:

$$p(x|q) \rightarrow p(x|q,a)$$

Stability of Pronunciation Models

- When decoding a lexical entity through a “perturbed” HMM topology, how much/fast does the inferred phonetic transcription change?
- In our case:
  - Perturbation: constrained -> unconstrained (relaxed) lexical model
  - Stability measured in terms of:
    - Confidence measure
    - Levenshtein distance wrt baseform

Relaxing Lexical Constraint



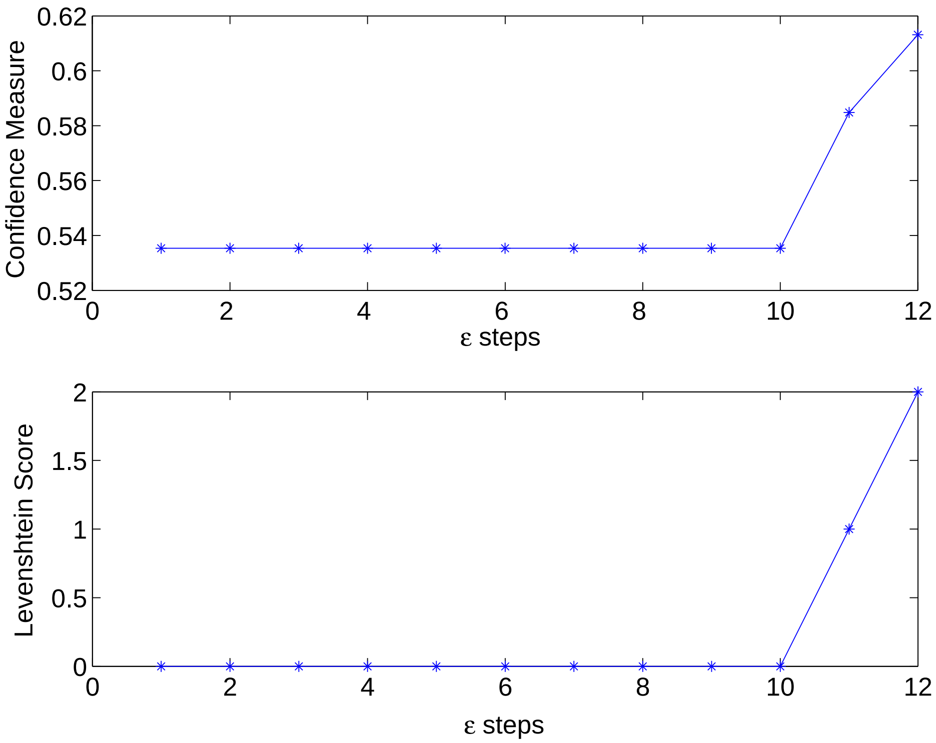
$\epsilon = \infty$, bi-gram phoneme language model

$\epsilon = 0$, unigram phoneme language model

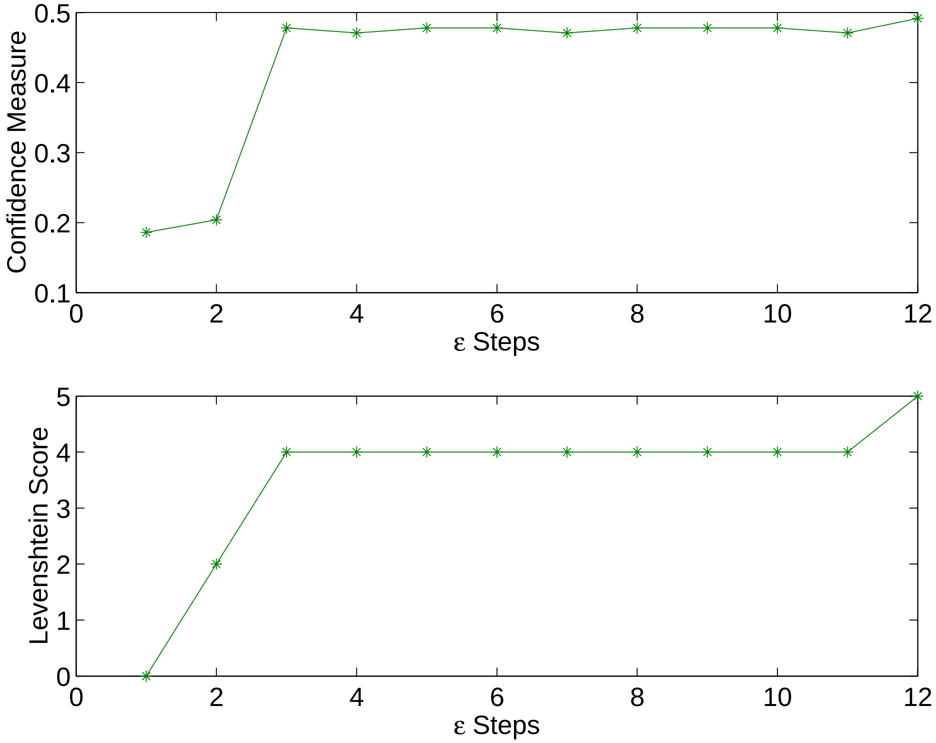
$$\begin{bmatrix} 0.0 & \frac{1}{3+3\epsilon} & \frac{1+3\epsilon}{3+3\epsilon} & \frac{1}{3+3\epsilon} & 0.0 \\ 0.0 & \frac{1}{4+4\epsilon} & \frac{1+4\epsilon}{4+4\epsilon} & \frac{1}{4+4\epsilon} & \frac{1}{4+4\epsilon} \\ 0.0 & \frac{1+4\epsilon}{4+8\epsilon} & \frac{1}{4+8\epsilon} & \frac{1}{4+8\epsilon} & \frac{1+4\epsilon}{4+8\epsilon} \\ 0.0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}$$

Stability of Pronunciation Models

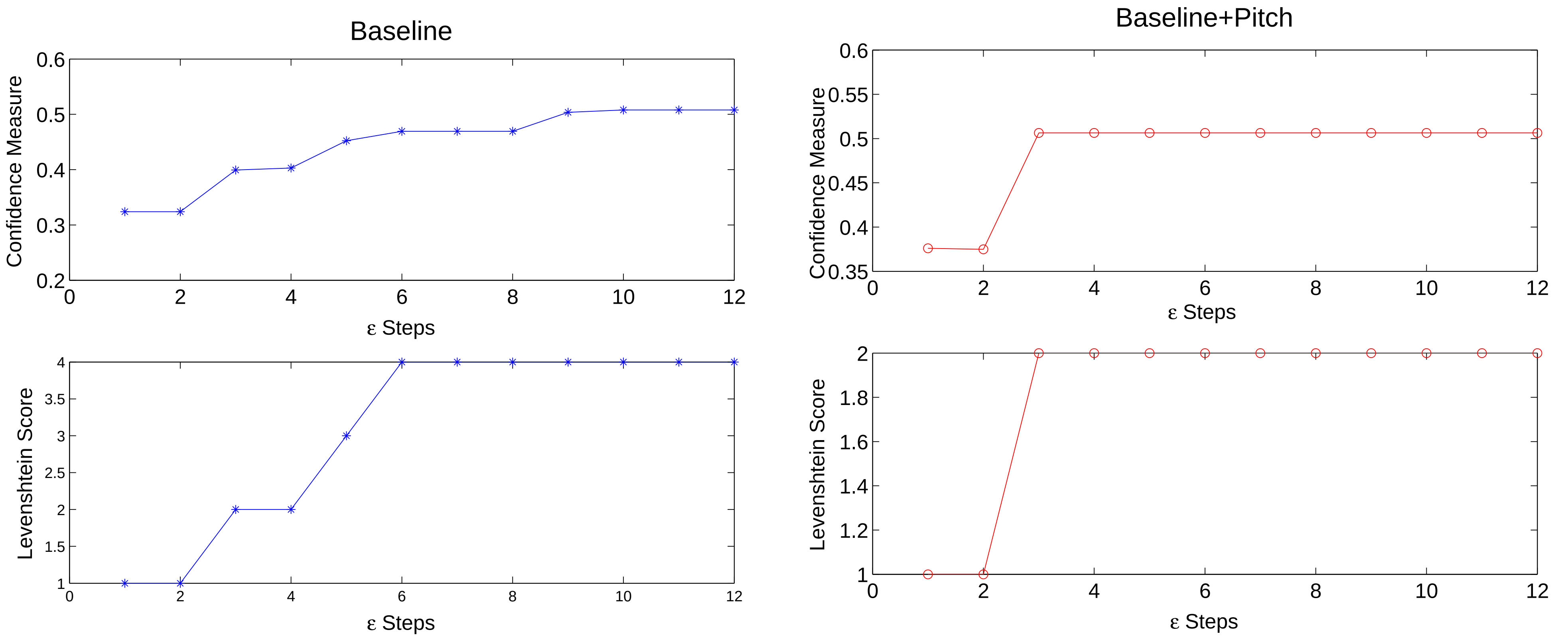
Stable



Unstable



Cont. (with Auxiliary Variable)



Experimental Setup

- Phonebook: Speaker-independent task-independent isolated word recognition.
- Vocabulary: 8 different sets of 75 word lexicon or single lexicon of 602 words.
  Number of context-independent phonemes: 42
- Acoustic feature: 21 dim. MFCC and ΔMFCC features.
  Auxiliary feature: pitch frequency and short-term energy.
- Training set: 19420 utt.; Validation set: 7290 utt.
  Development set: 2969 utt.; Test set: 3639 utt.

Lexical Models

Acoustic model: baseline+pitch

$\#models$ x $\#words$
1 x 441
2 x 106
3 x 48
4 x 7
Total words: 602
Total lexical forms: 825

Results

75 word lexicon (word error rate, expressed in %)

Systems	Original lexicon	Updated lexicon
baseline	4.2	3.0[†]
baseline+pitch	2.5	1.7[†]

602 word lexicon

Systems	Original lexicon (602)	Updated lexicon (825)
baseline	11.0	10.1
baseline+pitch	7.3	6.4

Conclusion and Future Work

- Preliminary studies yield significant performance improvement with limited number of lexical models.
- Can be used to evaluate and compare different acoustic models without recognition.
- To be extended to spontaneous speech recognition tasks.

Thank you for your attention!