Pronunciation Lexicon Development for Under-Resourced Languages Using Automatically Derived Subword Units: A Case Study on Scottish Gaelic

Marzieh Razavi, Ramya Rasipuram and Mathew Magimai. Doss

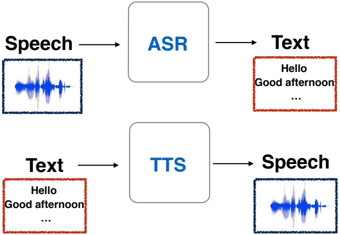
28 November 2015

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

RESEARCH INSTITUTE

Motivation

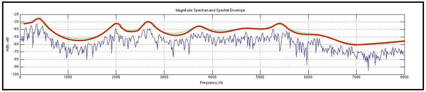
Speech Technology Systems



- Standard speech technology systems model words as a sequence of subword units.
- Using subword units necessitates availability of two resources:
  1. The subword unit set.
     - $\mathcal{A} = \{a^1, \cdots, a^d, \cdots, a^D\}$
  2. The Lexicon mapping each word to a sequence of subword units.

Lexical Resources

- The most commonly used subword units: **Phones**
  - Linguistically motivated units: /f/, /p/, /b/, $\cdots$
  - Spectral envelope depicts the characteristics of phones.



/lh/

- The phonetic lexicon provides the phonetic representation of words.

$$\text{phone} \rightarrow \text{/f/ /ow/ /n/}$$
$$\text{map} \rightarrow \text{/m/ /ae/ /p/}$$
$$\ldots$$

  - Typically developed manually.
  - Augmented using grapheme-to-phoneme (G2P) conversion approaches.
  - Require linguistic knowledge & human expertise.

Lexicon Development for Under-Resourced languages

- Majority languages have well-developed lexicons.
- Under-resourced languages may lack proper lexical resources.
  - Examples: Uspanteko, Haitian Creole, . . .
  - Linguistic knowledge and human expertise may be very limited.
    $\Rightarrow$ Conventional approaches cannot be exploited.

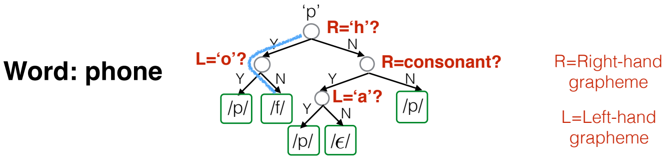
A possible solution

Automatically derive “phone-like” subword units and generate associated pronunciations using transcribed speech data.

Background

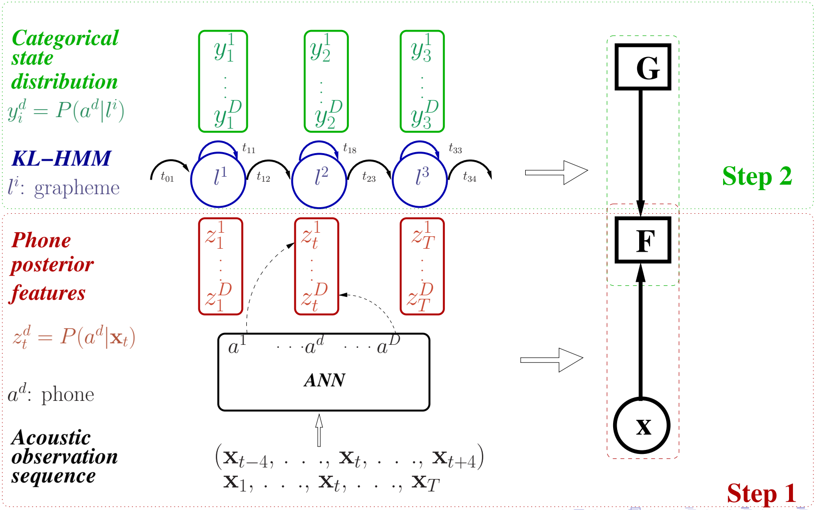
G2P Conversion

- Conventional Data-Driven G2P approaches:
  - Assumes the availability of a *seed lexicon* obtained using linguistic knowledge and expertise.
  - Apply machine learning techniques to learn the G2P relationship.

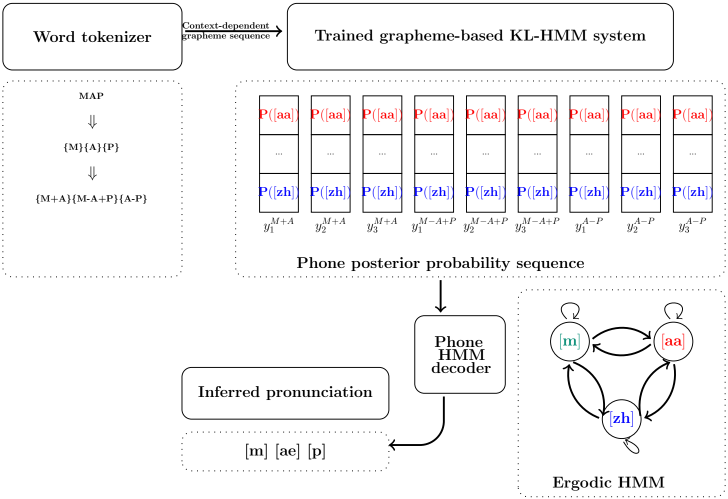


- Acoustic data-driven G2P conversion approach:
  - Assumes some speech data in addition to the seed lexicon is available.
  - G2P relationship is learned through acoustics.

Acoustic G2P Conversion Approach (I): Learning the G2P Relationship Using Acoustics

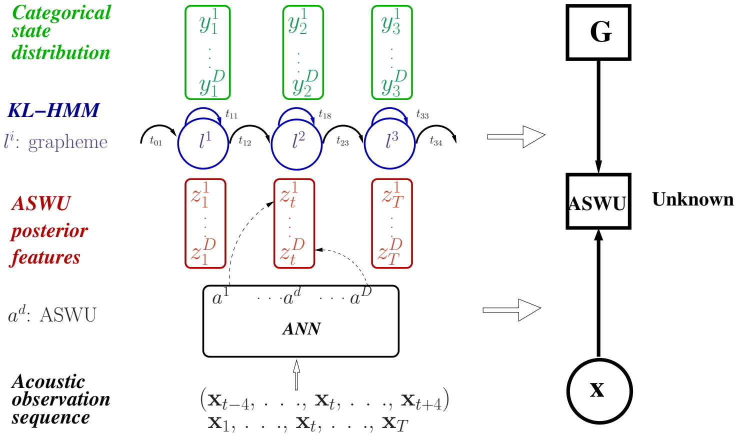


Acoustic G2P Conversion Approach (II): Pronunciation Inference Given the Learned G2P Relation



HMM-Based Formulation

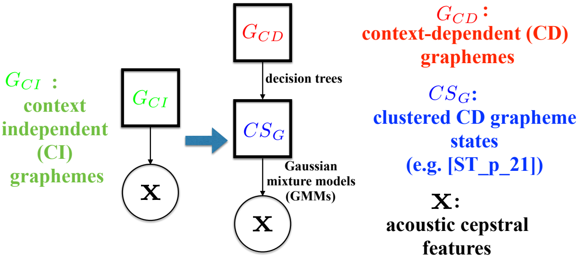
Formulating the Problem



- $\{a^d\}_{d=1}^D$ are automatic subword units (ASWUs) and unknown.
- Once $\{a^d\}_{d=1}^D$ discovered, apply the acoustic G2P conversion approach.

Standard HMM-based ASR

- Context-dependent (CD) subword units are modeled with HMMs with mixture of Gaussian state-output distributions.
- e.g. iphone $\rightarrow$ [i] [p] [h] [o] [n] [e] $\rightarrow$ [i+p] [i-p+h] [p-h+o] [h-o+n] [o-n+e] [n-e]



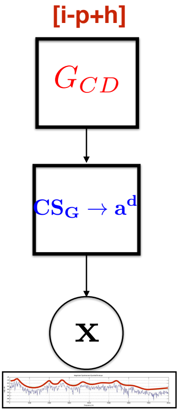
- Data sparsity issue
- Clustering and tying (parameter sharing) using decision trees:
  - maps each G_{CD} to a CS_G

Derivation of Automatic Phone-Like Subword Units

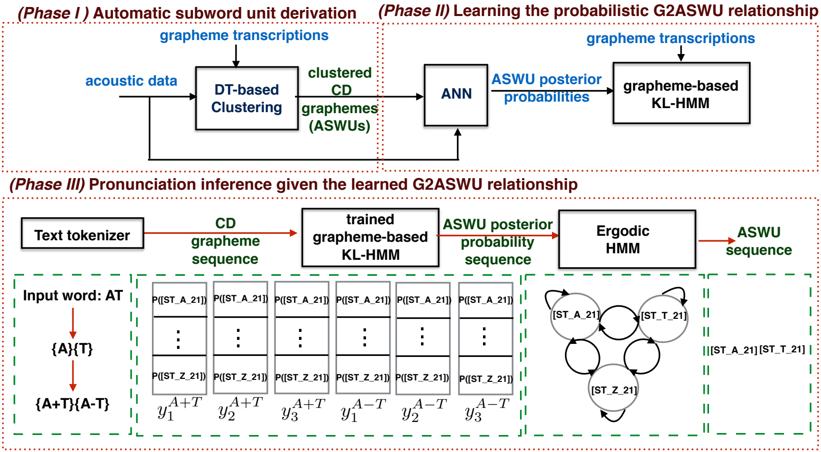
Hypothesis: The clustered CD grapheme states CS_G can be treated as phone-like automatic subword units a^d.

- Cepstral feature $\mathbf{x}$ carries phone-like information.
- G_{CD} tends to relate to a phone in a regular manner.
  - Example:
    G_{CI} [p] $\rightarrow$ /p/ , /f/
    G_{CD} [p+h] $\rightarrow$ /f/
- CS_G is found by maximizing the likelihood of the data.
- CS_G relates to both $\mathbf{x}$ and G_{CD} .

$\Rightarrow$ CS_G should be phone-like.



Block Diagram of ASWU Derivation and Pronunciation Generation



Experimental Studies

Scottish Gaelic

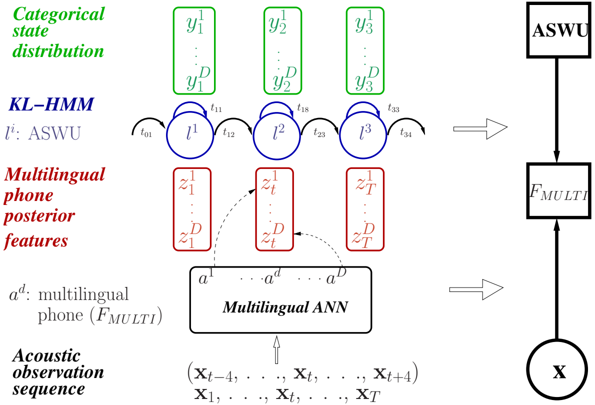
- Low-resourced and minority language; Endangered with only 60,000 speakers.
- Alphabet has 18 letters, consisting of 5 vowels and 13 consonants.
- There are 12 basic consonant types in Scottish Gaelic:
  - fortis or lenis: a grapheme [h] next to the consonant.
  - broad or slender : surrounded by ([a], [o], [u]) or ([e], [i])
- Number of graphemes is greater than number of phonemes in word.
  - an-diugh $\rightarrow$ /ə/ /n/ /d^j/ /u/
  - aghaidh $\rightarrow$ /γ : ./ /I/

Under-Resourced Language Study: Scottish Gaelic

- Corpus was collected by the University of Edinburgh.
- Recordings from broadcast news and discussion program.
- Use transcribed speech data for subword unit derivation and pronunciation generation.

Corpus	Lexicon size (in words)	# of grapheme subword units	Training data	Test data
Scottish Gaelic	5082	31	180 (min) (22 speakers)	60 (min) (12 speakers)

Relating ASWUs to Phonetic Units



- No phonetic lexicon available.
- Exploit auxiliary linguistic resources from other languages.

Analysis of ASWU-based Pronunciations

- Map each ASWU to most probable multilingual phone.
- Provide the ‘perceived’ pronunciations for each word through informal hearing.

Word	*Lex-ASWU-82*	mapped pron.	perceived pron.
MHÀL	[ST_B_22] [ST_À_21] [S_L_23]	/v/ /a/ /l/	/v/ /a/ /l/
THOG	[ST_T_21] [ST_O_23] [ST_G_23]	/h/ /o/ /k/	/h/ /O/ /g/
PHÒS	[ST_F_21] [ST_Ò_21] [ST_S_23]	/f/ /o/ /s/	/f/ /o/ /s/

- The ASWU-based pronunciations to a certain extent capture the linguistic rules related to pronunciations.

ASR Study

System	# of units	# of tied states	WA
HMM-GMM-GRAPH	32	1158	64.6
HMM-GMM-ASWU	82	1161	**66.4**

- HMM-GMM-GRAPH : Grapheme-based ASR
- HMM-GMM-ASWU : Proposed approach
- The ASWU-based lexicon yields a significantly better ASR system than the grapheme-based lexicon.

Summary

- Proposed an HMM formalism to derive phone-like subword units and generate associated pronunciations.
- The formalism is scalable to under-resourced languages.
- Investigated the potential of ASWUs for developing linguistically meaningful lexicons.
- Interpreted ASWUs in terms of linguistic units by exploiting auxiliary languages resources and prior linguistic knowledge.

Thank You