Acoustic Unit Discovery and Pronunciation Generation from a Grapheme-Based Lexicon

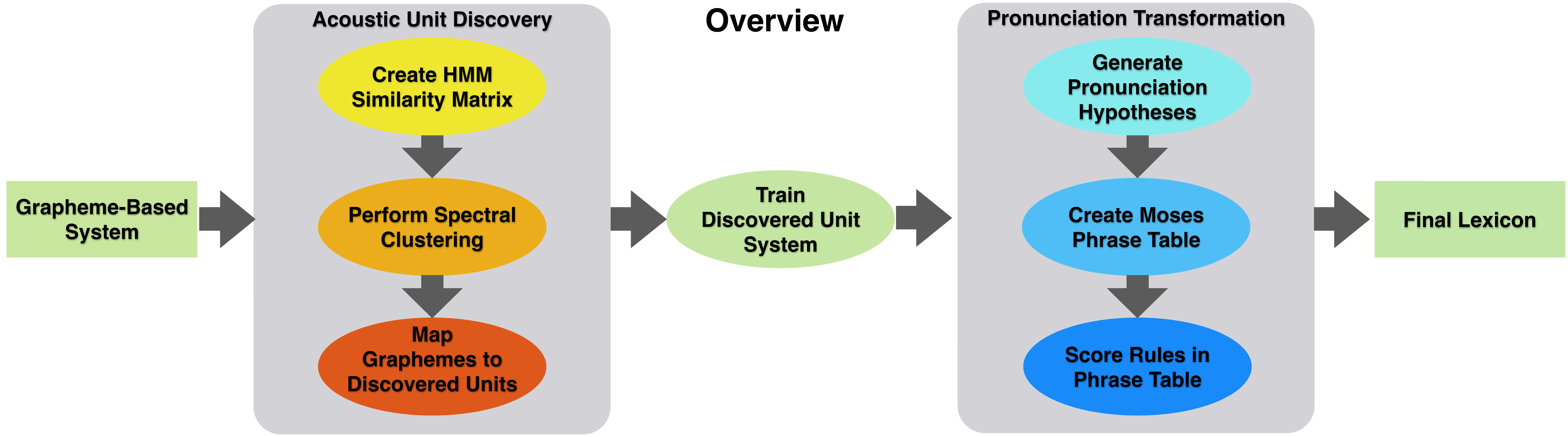
William Hartmann, Anindya Roy, Lori Lamel, and Jean-Luc Gauvain

Spoken Language Processing Group, LIMSI-CNRS {hartmann, roy, lamel, gauvain}@limsi.fr

Introduction

- Unlike the other main components of an ASR system, the pronunciation lexicon is largely handmade.
- Low-resource languages may not have expert-defined lexicons.
- We propose a two-stage approach to learning both the lexicon and the underlying acoustic units.
- Our approach relies on an initial baseline grapheme-based system.
- Acoustic units are learned by clustering the context-dependent grapheme-based models.
- Pronunciations are generated by transforming the original lexicon with an SMT-based approach.
- Each individual stage produces a significant improvement over the baseline system.
- Combined, the approach reduces the relative word error rate by 13%.

Overview



Acoustic Unit Discovery

- Acoustic units are discovered by clustering context-dependent grapheme-based HMMs.
- Requires defining a similarity measure between individual HMMs (Equation 1).
- CSD is the Cauchy-Schwarz Divergence measure (Equation 2).
- We use the CSD because a closed form solution for a Mixture of Gaussians exists.

$$\mathrm{HMM_{sim}}(\mathbf{h},\mathbf{h}') = \sum_{a=1}^{A}\sum_{b=1}^{B}\frac{\alpha_{a,b}}{\mathrm{CSD}(h_a, h'_b)+1}$$

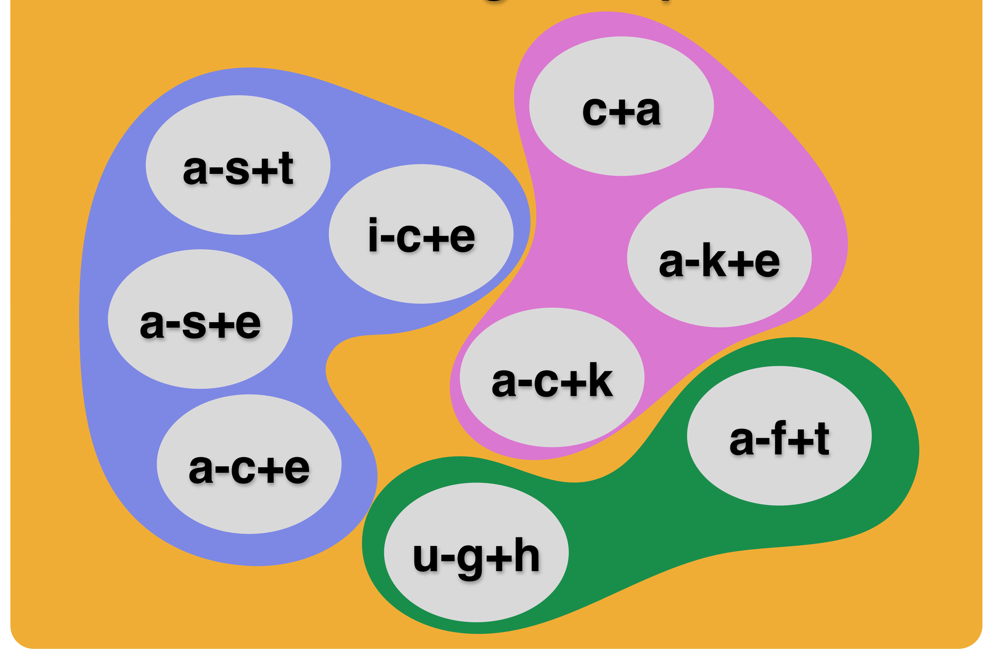
Equation 1

$$\mathrm{CSD}(\mathbf{p},\mathbf{q}) = -\log\frac{\sum_i p_i q_i}{\sqrt{\sum_i p_i^2 \sum_i q_i^2}}$$

Equation 2

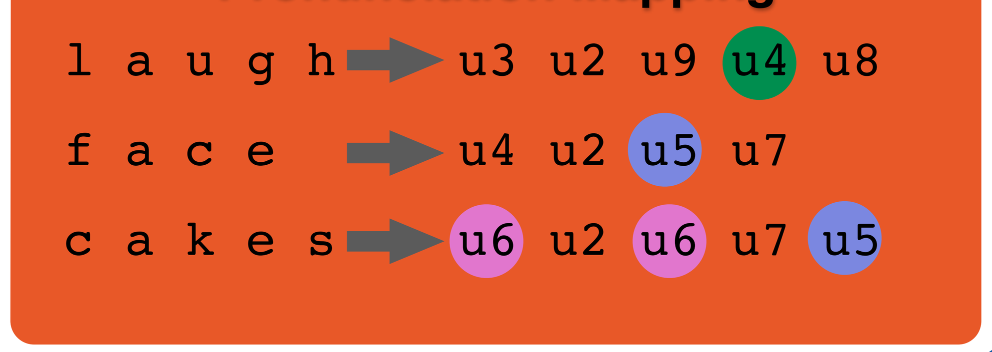
- Clustering is performed using spectral clustering.
- We achieved better performance with a k-nearest neighbor similarity graph rather than a similarity matrix.
- Since the optimal number of acoustic units is not known a priori, we tried various numbers of clusters.

Clustering Example



- Based on the clustering, pronunciations are mapped to the new acoustic units.
- Each pronunciation will have the same number of units as in the baseline grapheme-based lexicon.

Pronunciation Mapping



Pronunciation Transformation

- Context-dependent acoustic models are trained.
- The training data is decoded in terms of the acoustic units.
- Based on the time-aligned results, each word in the training set has one or more pronunciation hypotheses.
- The example shown uses grapheme units for clarity.

lack	l a k
lack	l e k
lochs	l o c x s
necessary	n e s e s r y
ford	f r d
ford	f r n
caught	k o t

- Using Moses, a phrase table is learned from the pronunciation hypotheses.
- Each entry in the table represents the translation of a sequence of acoustic units into an alternate sequence.
- Moses also allows for reordering of units, but we disable this option for our work.
- The phrase table can transform the pronunciations in the original lexicon.

Source	Target	$p(t \mid s)$
a c k	a k	0.19
c h s	c x s	0.13
c e s s a	s e s e	0.36
f o r d	f r d	0.17
a u g h t	o t	0.25

- Using the translation table directly would decrease performance, so we first prune the table.
- Each rule is scored individually.
- We measure the average change in likelihood when aligning the training data after transformation.
- Only rules that surpass a certain threshold are kept.
- The final transformation works for words both seen and unseen during training.

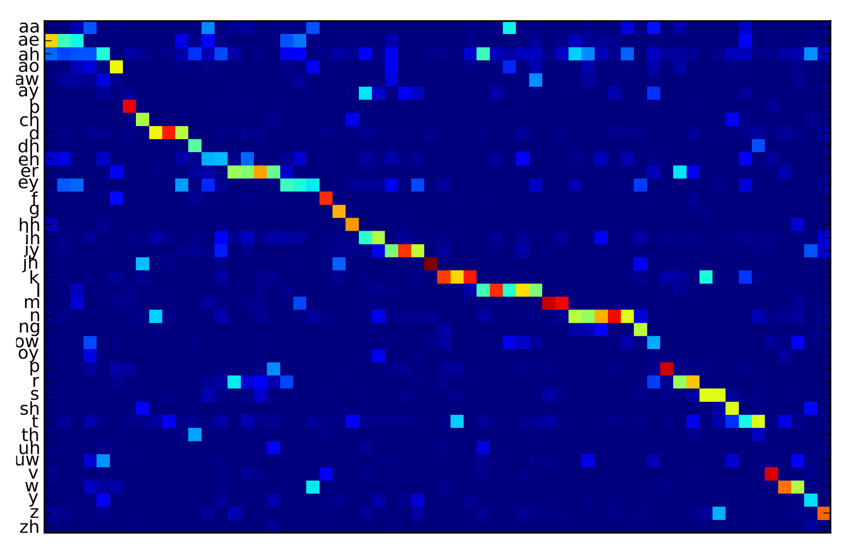
Source	Target	Δ LLH
a c k	a k	-13.05
c h s	c x s	48.25
c e s s a	s e s e	63.28
f o r d	f r d	-81.47
a u g h t	o t	87.39

Results

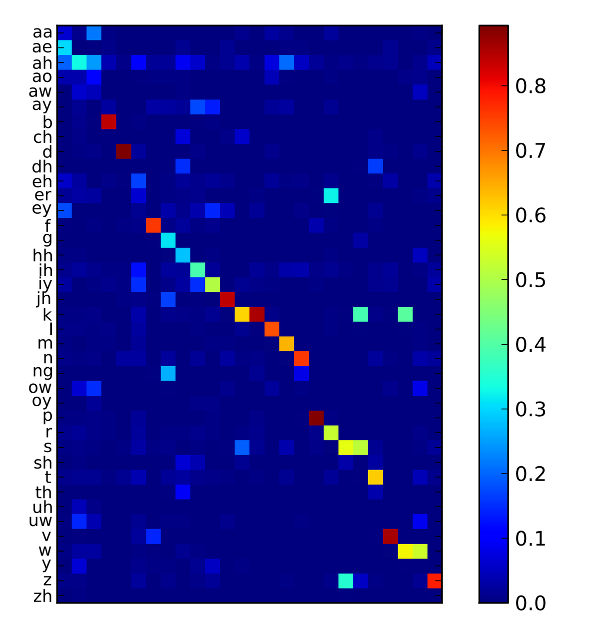
- WER results are presented on the WSJ0 5k word task.
- Acoustic models are trained with HTK.
- All recognizers use context-dependent models with 2000 tied states.

Unit Type	# Units	Direct	Trans.
Grapheme	26	15.8	14.5
Discovered	39	15.0	13.9
Discovered	50	15.2	13.9
Discovered	60	**14.4**	**13.8**

- Decoding is performed with a bigram LM.
- Direct refers to the lexicon after mapping, and Trans. refers to the lexicon after pronunciation transformation.
- The figures below demonstrate the correlation between the discovered units and phones, and the grapheme units and phones.



Discovered 60



Grapheme

- We have proposed a two-stage approach for acoustic unit discovery and pronunciation generation that reduces relative WER by 13% compared to a baseline grapheme-based system.
- We are currently working to apply these techniques to other languages.

This research was partially supported by OSEO, the French State agency for innovation, under the Quaero program and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.