Integrating Thai Grapheme Based Acoustic Models into the ML-Mix Framework - For Language Independent and Cross-Language ASR

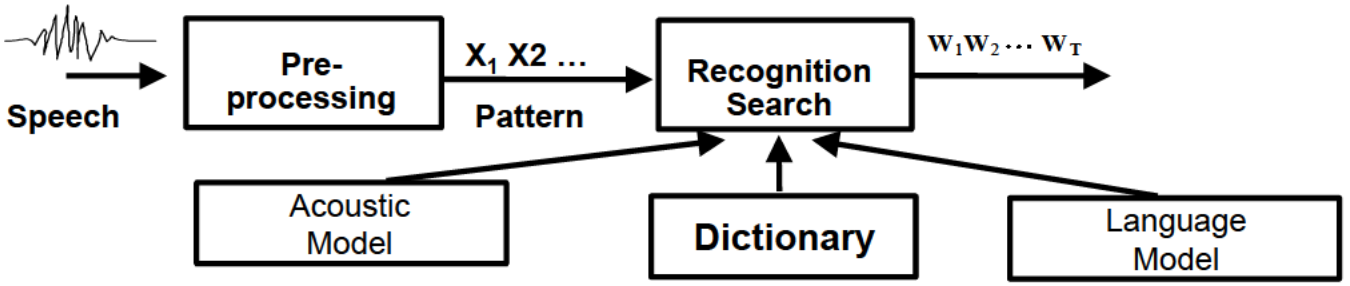
Sebastian Stüker

SLTU Workshop

Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825

Grapheme Based ASR

- Traditionally ASR Systems use phonemes or sub-phonemes as modeling units
- Requires a pronunciation dictionary for mapping orthography of words to model sequence
- Creation of a dictionary can be costly in terms of time and money
- Dictionaries can be learned automatically, but that requires training material and often manual post-processing
- For less prevalent languages it can be difficult to find the necessary expert or to find the necessary resources
- Using Graphemes instead of phonemes as modeling units makes the creation of a pronunciation dictionary a trivial task



Grapheme Based ASR

- Has been shown to work for many languages, e.g: German, English, Spanish, Russian, Arabic, Thai
- Terminology changes accordingly: Sub-Graphemes, Polygraphemes, Grapheme Decision Tree
- When building context-dependent model, the use of classification and regression tree for clustering the polyphone models is common
- Polyphone Decision Trees (PDTs) usually use linguistically motivated questions, such articulatory properties of the phonemes
- For graphemes these properties do not exist; questions should not require any phonetic knowledge
- Previous work has shown that asking for the identity of the graphemes in the context of a polygrapheme works very well; requires no knowledge
- Polygrapheme Tree captures implicitly the relationship between graphemes and acoustic realisation

Corpus and Task

- GlobalPhone: dictated newspaper texts (like WSJ0)
- Collected in an uniform way across many languages (currently 19). Well suited for multilingual ASR and porting research
- Languages selected for this work: English (EN), Russian (RU), Spanish (SP), and Thai (TH)
- For porting experiments EN, RU, and SP take the role of well-known languages; Thai takes the role of the new language about which little is known
- For EN, RU, and SP sufficient training material is available for Thai only little adaptation data. Training (train), development (dev), and evaluation (eval) sets availabel for all.

	EN	RU	SP	TH
train	15	17	17.6	24.5
dev	0.4	1.3	2.1	1.3
eval	0.4	1.6	1.7	1.1
adapt	---	---	---	0.5

Baseline Systems

- Baseline Systems for comparing word error rates
  - Trained with the help of forced alignments from previous systems
  - Standard MFCC based pre-processing
  - Standard EM training
- For Thai no pre-processing of the graphemes was done (we assume no extensive knowledge)
- Differences in the Word Accuracies reflect suitability for grapheme based approach and language inherent differences

		EN	RU	SP	TH
CI	dev	45.8%	48.1%	55.7%	70.8
	eval	46.5%	44.2%	68.6%	71.3
CD	dev	84.4%	64.3%	78.0%	87.3
	eval	82.7%	60.7%	85.9%	86.0

Multilingual ASR using ML-Mix

- Assume that phonemes are pronounced the same across languages
- “Multilingual” refers to systems that can recognize multiple languages seen during training, using only one acoustic model
- ML-Mix
  - Phonemes that are common to multiple languages share the training material from that language
  - Information about to which language a phoneme belongs to is discarded
- ML-Mix models can also be used for bootstrapping acoustic models in new languages or to recognize languages which have not been seen during training
- The same principles can be applied to grapheme based models
  - Sharing of models can be worse depending on the languages involved
  - Assumption that graphemes are pronounced the same across languages does not hold
  - Therefore worse performance than for phonemes

ML3-Mix

- Grapheme Based System trained on the languages EN, RU, and SP
- For RU we assume a romanized transcription as given, so that a partial sharing of models is possible
- Incorporation of TH not possible, nor test on TH, because no mapping from ML3-Mix graphemes to TH graphemes

		EN	RU	SP
ML-3Mix-CI	dev	27.6%	38.5%	44.5%
	eval	29.2%	33.8%	58.5%
ML-3Mix-CD	dev	78.2%	60.5%	74.7%
	eval	75.9%	44.2%	83.7%

ML3-Mix word accuracies on the training languages

Distance Measures between models

- Goal: Find a data driven mapping from ML3-Mix models to the Thai graphemes
- From literature known multiple distance measures between Gaussian mixture models
- Trained 1 Gaussian per model helper models for ML3-Mix and for Thai on the Thai adapt material
- Establish mapping by minimizing distance between models

Gaussian based distance measures

- Euclidean Distance between mean vectors
  - Does not consider variances of the models

$$d_{eucl}(\Gamma_1, \Gamma_2) = \sqrt{(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T}$$

- Extended Mahalanobis Distance
  - Considers variances of the models

$$d_{extMhn}(\Gamma_1, \Gamma_2) = \sqrt{(\mu_1 - \mu_2)^T (\Sigma_1 + \Sigma_2)^{-1} (\mu_1 - \mu_2)}$$

Gaussian based distance measures

Kullback Leibler Distance

- Based on Kullback Leibler Divergence
- Problem: Not symmetric, does not obey triangle inequality
- Make it symmetric by summing KL divergences in both “directions”

$$d_{kl}(P_1, P_2) = \int P_1(x) \log \frac{P_1(x)}{P_2(x)}$$

$$d_{kl-sym}(\Gamma_1, \Gamma_2) = d_{kl}(\Gamma_1, \Gamma_2) + d_{kl}(\Gamma_2, \Gamma_1)$$

$$d_{kl-sym}(\Gamma_1, \Gamma_2) = \frac{1}{2} \sum_{i=1}^{d} \frac{\sigma_{1,i}^2}{\sigma_{2,i}^2} + \frac{\sigma_{2,i}^2}{\sigma_{1,i}^2} - 2 + \left(\frac{1}{\sigma_{1,i}^2} + \frac{1}{\sigma_{2,i}^2} \right) (\mu_{1,i} - \mu_{2,i})^2$$

Gaussian based distance measures

- Bhattacharya Distance
  - Often used in two class scenario
  - Symmetric, but does not obey triangle inequation

$$d_{bhatt}(P_1, P_2) = -\ln\left(\int_x \sqrt{P_1(x)P_2(x)}\right)$$

$$d_{bhatt}(\Gamma_1, \Gamma_2) = \frac{1}{2}\sum_{i=1}^{d} \ln\left(\frac{\sigma_{1,i}^2 + \sigma_{2,i}^2}{2\sqrt{\sigma_{1,i}^2\sigma_{2,i}^2}}\right) + \frac{|\mu_{1,i} - \mu_{2,i}|^2}{2\left(\sigma_{1,i}^2 + \sigma_{2,i}^2\right)}$$

Applying ML3-Mix to Thai

- Map context-independent models to closest Thai helper model
- Only works for context-independent models

	dev	eval
Euclidean	13.9%	15.9%
Ext. Mahalanobis	16.7%	16.9%
Kullback-Leibler	20.8%	19.7%
Bhattacharya	20.8%	19.0%

WA when applying CI ML3-Mix to Thai using various distance measures

- Gain from Euclidean to other measures shows that variance contains valuable information
- Kullback-Leibler and Bhattacharya outperform extended Mahalanobis distance

Incorporating Thai into the ML-Mix Framework

- Same procedure for mapping the Thai training data to the models in the ML3-Mix model
- Bhattacharya distance turns out to be the best
- Trained CI and CD models on all training material

		EN	RU	SP	TH
ML-4Mix-CI	dev	22.1%	32.5%	38.3%	44.1%
	eval	22.7%	28.1%	50.7%	44.6%
ML-4Mix-CD	dev	73.5%	57.9%	71.9%	68.3%
	eval	72.2%	54.3%	81.5%	68.3%

Conclusion

- Studied data driven mappings between Latin alphabet based, grapheme ASR acoustic models to Thai grapheme based models
- Used the approach for cross-language application on Thai and incorporating Thai into a grapheme based ML-Mix model
- Future work will study porting context dependent models to Thai using the data driven distance measures