Pairwise Learning using Multi-lingual Bottleneck Features for Low-resource Query-by-example Spoken Term Detection

Yougen Yuan[1,2], Cheung-Chi Leung[2], Lei Xie[1], Hongjie Chen[1], Bin Ma[2], Haizhou Li[2,3]

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Institute for Infocomm Research, A*STAR, Singapore
[3]Department of ECE, National University of Singapore, Singapore

March 9, 2017

Outline

Outline

1. Introduction
   - Background
   - Motivation and contribution
2. Methods
   - Multi-lingual bottleneck features (BNFs)
   - Pairwise learning
   - Query-by-example spoken term detection (QbE-STD)
3. Experiments
   - Data and evaluation
   - Results and analysis
4. Conclusions

Problem description of low-resource query-by-example spoken term detection (QbE-STD)

- A search problem for the occurrence of a spoken query in audio archives.
- Limited training data in low-resource scenarios.
- Difficult to give utterances with labels if no prior linguistic knowledge in the language.

Previous work

- Extract unsupervised acoustic features directly in low-resource target languages [1, 2, 3, 4].
- Extract posterior or bottleneck features (BNFs) from neural networks (NNs) trained using high-resource non-target languages [5, 6, 7, 8, 9].

Outline

1. Introduction
   - Background
   - Motivation and contribution
2. Methods
   - Multi-lingual bottleneck features (BNFs)
   - Pairwise learning
   - Query-by-example spoken term detection (QbE-STD)
3. Experiments
   - Data and evaluation
   - Results and analysis
4. Conclusions

Motivation

- Pairwise learning
  - Training NNs with paired examples.
  - Successful for various tasks, including face verification [10], sentence similarity [12], phone discrimination [11], and our previous study [13] on a word discrimination task.
- Multi-lingual BNFs
  - A kind of compact representations.
  - More language-independent and more flexible for rapid language adaptation; especially in low-resource languages.

Motivation

- Pairwise learning
  - Training NNs with paired examples.
  - Successful for various tasks, including face verification [10], sentence similarity [12], phone discrimination [11], and our previous study [13] on a word discrimination task.
- Multi-lingual BNFs
  - A kind of compact representations.
  - More language-independent and more flexible for rapid language adaptation; especially in low-resource languages.

Contribution

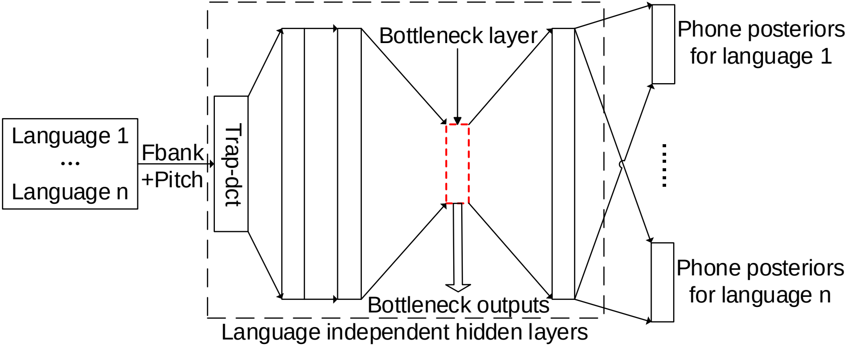
- The first attempt to use pairwise learning based on multi-lingual BNFs.
- The first attempt to use pairwise learning for QbE-STD.

Outline

1. Introduction
   - Background
   - Motivation and contribution
2. Methods
   - Multi-lingual bottleneck features (BNFs)
   - Pairwise learning
   - Query-by-example spoken term detection (QbE-STD)
3. Experiments
   - Data and evaluation
   - Results and analysis
4. Conclusions

Multi-lingual BNF extraction

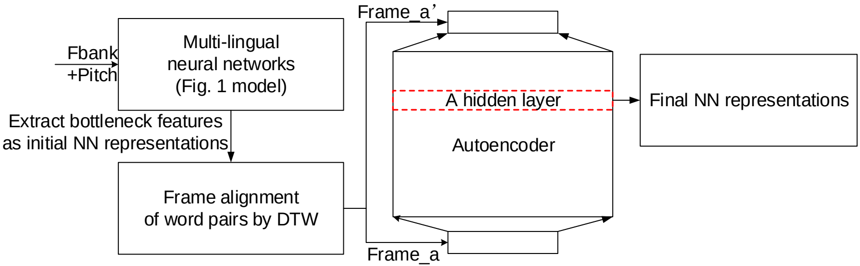
- Train a multi-lingual bottle-type NN from non-target languages.



Outline

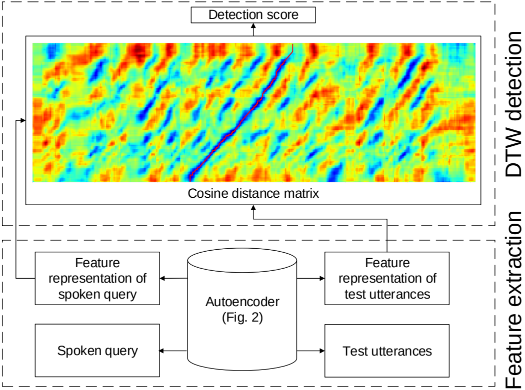
Pairwise learning with an autoencoder

- Align two sequences of multi-lingual BNFs with DTW.
- Train a pre-trained AE with Mean Squared Error (MSE) using aligned frame pairs.
- Extract newly learned feature representation from an internal hidden layer of trained NN.



Outline

NN-based template matching method for QbE-STD



Outline

1. Introduction
   - Background
   - Motivation and contribution
2. Methods
   - Multi-lingual bottleneck features (BNFs)
   - Pairwise learning
   - Query-by-example spoken term detection (QbE-STD)
3. Experiments
   - Data and evaluation
   - Results and analysis
4. Conclusions

Data

- Target language (for QbE-STD)

Corpus	Training set (No. of word pairs)	Keyword set (No. of examples)	Test set (No. of utterances)
TIMIT [3, 4]	10,000	346	944
Switchboard	100,000	346	100

- Non-target languages (for multi-lingual BNFs extractor)
  - HKUST Mandarin Chinese (LDC2005S15; 170hr)
  - Fisher Spanish (LDC2001S01; 152hr)

Data

- Target language (for QbE-STD)

Corpus	Training set (No. of word pairs)	Keyword set (No. of examples)	Test set (No. of utterances)
TIMIT [3, 4]	10,000	346	944
Switchboard	100,000	346	100

- Non-target languages (for multi-lingual BNFs extractor)
  - HKUST Mandarin Chinese (LDC2005S15; 170hr)
  - Fisher Spanish (LDC2001S01; 152hr)

Metrics of evaluation

MAP : the mean average precision of each query in the test set.

P@N : the average precision of the top N utterances where N is the number of the correct hit utterances in test set.

P@5/P@10 : the average precision of the first five or ten ranked utterances.

Outline

QbE-STD on TIMIT and Switchboard

Corpus	Representation	No pairwise training (MAP/P@N/P@10)	Pairwise training (MAP/P@N/P@10)
TIMIT	MFCCs	0.285/0.289/0.247	0.297/0.293/0.257
	BNFs (Mandarin)	0.494/0.459/0.413	0.571/0.538/0.467
	BNFs (Spanish)	0.540/0.512/0.446	**0.594/0.561**/0.484
	BNFs (Multi-lingual)	**0.552/0.524/0.461**	**0.594/0.561/0.490**
Switchboard	MFCCs	0.232/0.200/0.232	0.258/0.236/0.260
	BNFs (Mandarin)	0.370/0.338/0.446	0.417/0.382/0.451
	BNFs (Spanish)	0.388/0.358/0.475	0.430/0.398/**0.484**
	BNFs (Multi-lingual)	**0.400/0.365/0.485**	**0.435/0.404**/0.473

Analysis

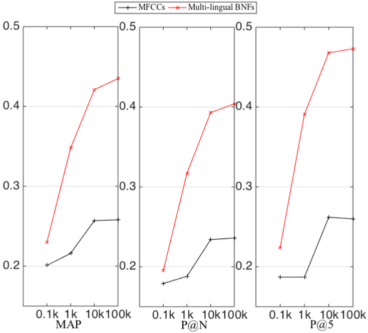
- Multi-lingual BNFs
  - Are much better than MFCCs.
  - Usually outperform the cross-lingual BNFs.
- Pairwise learning
  - Provides a more efficient feature representation for QbE-STD.
  - Usually hold the best performance with multi-lingual BNFs in the QbE-STD tasks.

Analysis

- Multi-lingual BNFs
  - Are much better than MFCCs.
  - Usually outperform the cross-lingual BNFs.
- Pairwise learning
  - Provides a more efficient feature representation for QbE-STD.
  - Usually hold the best performance with multi-lingual BNFs in the QbE-STD tasks.

Dependence on the amount of word-pair supervision

- With more word pairs, pairwise learned NN feature representation gives a better performance.
- With 10,000 word pairs, pairwise learned features give comparable performance to those using all word pairs.



Effect of input features and frame alignment

- Regardless of either MFCCs or multi-lingual BNFs are used for frame-level DTW alignment, multi-lingual BNFs consistently provide much better QbE-STD results than MFCCs as input features.

Corpus	Input features of AE	Features for alignment	
		MFCCs	BNFs (Multi-lingual)
TIMIT	MFCCs	0.285/0.289/0.247	0.320/0.314/0.274
	BNFs (Multi-lingual)	0.587/0.556/0.486	**0.594**/**0.561**/**0.490**
Switchboard	MFCCs	0.258/0.236/0.260	0.273/0.248/0.286
	BNFs (Multi-lingual)	0.432/0.395/**0.483**	**0.435**/**0.404**/0.473

Conclusions and future work

- We have proposed to perform pairwise learning using multilingual BNFs of word pairs for QbE-STD.
- Pairwise learning makes the resulted features more capable in phonetic discrimination for a new target language.
  - Brings further performance improvement on low-resource QbE-STD tasks.
- In future work, we will investigate methods of word-level pairwise learning for this task, which avoids frame-level alignment of word pairs.

References I

[1] Yaodong Zhang and James R Glass. "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams". In: *Proc. ASRU*. 2009, pp. 398–403.

[2] Haipeng Wang et al. "An acoustic segment modeling approach to query-by-example spoken term detection". In: *Proc. ICASSP*. 2012, pp. 5157–5160.

[3] Peng Yang et al. "Intrinsic spectral analysis based on temporal context features for query-by-example spoken term detection". In: *Proc. INTERSPEECH*. 2014, pp. 1722–1726.

[4] Hongjie Chen et al. "Unsupervised bottleneck features for low-resource query-by-example spoken term detection". In: *Proc. INTERSPEECH*. 2016, pp. 923–937.

References II

[5] Javier Tejedor et al. "Comparison of methods for language-dependent and language-independent query-by-example spoken term detection". In: *ACM Transactions on Information Systems* 30.3 (2012), p. 18.

[6] Luis J Rodriguez-Fuentes et al. "High-performance query-by-example spoken term detection on the SWS 2013 evaluation". In: *Proc. ICASSP*. 2014, pp. 7819–7823.

[7] Yang Peng et al. "The NNI query-by-example system for mediaeval 2014". In: *Proc. MediaEval Workshop*. 2014.

[8] Hou Jingyong et al. "The NNI query-by-example system for mediaeval 2015". In: *Proc. MediaEval Workshop*. 2015.

[9] Cheung-Chi Leung et al. "Toward high-performance language-independent query-by-example spoken term detection for mediaeval 2015: post-evaluation analysis". In: *Proc. INTERSPEECH*. 2016, pp. 3703–3707.

References III

[10] Sumit Chopra, Raia Hadsell, and Yann LeCun. "Learning a similarity metric discriminatively, with application to face verification". In: *Proc. CVPR*. 2005, pp. 539–546.

[11] Gabriel Synnaeve, Thomas Schatz, and Emmanuel Dupoux. "Phonetics embedding learning with side information". In: *Proc. SLT*. 2014, pp. 106–111.

[12] Jonas Mueller and Aditya Thyagarajan. "Siamese recurrent architectures for learning sentence similarity". In: *Proc. AAAI*. 2016, pp. 2786–2792.

[13] Yougen Yuan et al. "Learning neural network representation using cross-lingual bottleneck features with word-pair information". In: *Proc. INTERSPEECH*. 2016, pp. 788–792.