CROSS-LINGUAL SPEECH-BASED TOBI LABEL GENERATION USING BIDIRECTIONAL LSTM

Marco Vetter [1] *marco.vetter.mp8@is.naist.jp*
Sakriani Sakti [1,2] *ssakti@is.naist.jp*
Satoshi Nakamura [1,2] *s-nakamura@is.naist.jp*

[1] Nara Institute of Science and Technology, Japan
[2] RIKEN, Center for Advanced Intelligence Project AIP, Japan

Introduction

Motivation

- **Documenting under-resourced languages** automatically from speech requires word segmentation and lexical discovery
- Word segmentation is a difficult task
- **Infants learn prosody** before they learn words
- Prosodic information has also been experimentally shown to **improve word segmentation** [Ludusan, 2015]

Existing work

- Existing work on ToBI label prediction is usually monolingual [Chen, 2004] [Elvira-Garcia, 2016]
- Existing approaches use additional information at test time, not just speech [Rosenberg, 2010]

Problem

- Lack of training data on under-resourced languages

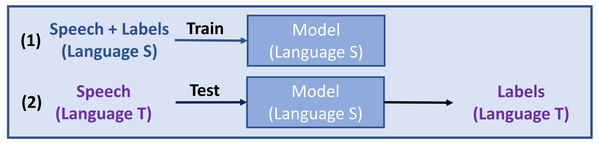
This work proposes

- We use bidirectional LSTM to **extract prosodic information from speech** in a cross-lingual fashion
- No labelled target language data is required

Methods

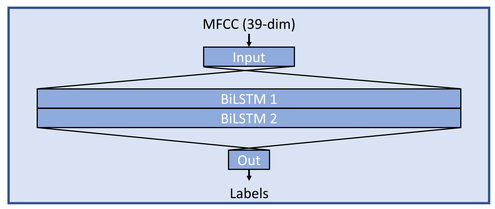
Cross-lingual model application

- **Cross-lingual** model application allows training on richly sourced languages (1)
- Models are then applied on an under-resourced target language (2)



Speech based label generation

- Our systems are entirely **speech-based** at test time
- No transcriptions, existing word segmentation or lexicon used



Cross-Lingual ToBI Break Index Labels

Corpora

- Corpus of spontaneous Japanese (~38h of spontaneous Japanese speech)
- Boston Radio Corpus (~78m of English radio speech)
- For this study we treat Japanese as a richly sourced language and English as a low-resource target language

Labelling system

- **ToBI break labels** are a standard available for various languages to mark prosody in speech.
- For cross-lingual prosody detection we map ToBI (EN) and J_ToBI (JP) labels to produce a common inventory

Break level description	ToBI	J_ToBI	Mapping
Word boundary	1	1	1
Lower-level grouping	2	n/a	2
Intermediate/accentual phrase	3	2	2
Intonational phrase	4	3	3

Network, Features and Metrics

Bidirectional LSTM using MFCC features

- The network is a bidirectional LSTM consisting of two hidden layers with 1024 BiLSTM cells each
- Features were standard 39-dimensional MFCC with Δ and ΔΔ

F-Score and tolerance

- Segmentation is evaluated using **Precision, Recall and F-score**
- As in other boundary detection tasks we apply a tolerance during evaluation (80ms)

Experimental Results

Binary and multi-class models

- Binary models only discriminate "break" or "no break"
- Multi-class models differentiate three label types (see "Data")

Results for binary labels

- These experiments discriminate only "break" and "no break"

Language pairing	P	R	F
JP → JP	0.4932	0.6137	0.5400
EN → EN	0.6308	0.7880	0.6956
JP → EN	0.5914	0.5216	0.5533

Experimental Results (cont.)

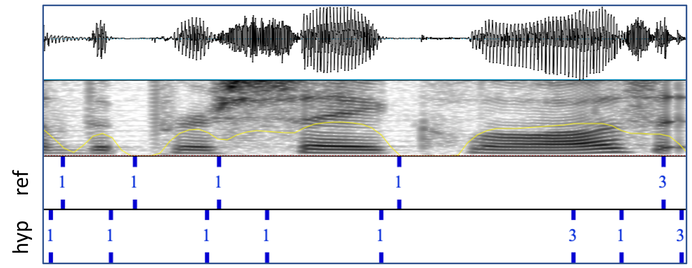
Results for multi-class labels

- These experiments discriminate between three classes
- (1) word boundary, (2) intermediate/accentual phrase, (3) intonational phrase

Language pairing	Class	P	R	F
JP → JP	1	0.4947	0.6272	0.5504
	2	0.3306	0.1620	0.2114
	3	0.4723	0.3770	0.3991
EN → EN	1	0.5238	0.3614	0.4177
	2	0.6330	0.0137	0.0225
	3	0.7031	0.1700	0.2555
JP → EN	1	0.3816	0.6578	0.4765
	2	0.1279	0.1299	0.1229
	3	0.2204	0.1974	0.2050

Visualized Output

- Cross-lingually generated boundary labels for English speech



- Ground truth labels are correctly identified, but placement does not match exactly
- False positives also occur with some frequency

Conclusions

Summary

- We **cross-lingually predict ToBI-style prosodic boundaries**
- Approach requires no prior knowledge of target language
- Uses no information besides speech at test time
- Models **retain much of their predictive power** when applied across languages.

Future Work

- Apply the extracted information to word segmentation
- Evaluate additional language pairs

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.