4C-03

日本語音声データを利用した韓国語音響モデルの学習

太刀岡勇気 †
デンソーアイティーラボラトリ †

1 はじめに

多言語音声認識を行う際に，音響モデルを複数の言語で同時に学習することがよく行われている [1]．大量の学習データを用意可能な補助言語を用いることで，少量データしか用いることのできない対象言語の音声認識性能を向上させることができるが，音素セットが言語によって異なる点が問題になりうる．そのため，国際音声記号 (IPA) などの汎用的な音素の集合を用いて分類器を構築する方式 [2] や，汎用文字符号化 [3] が使われる．しかしながら，このアプローチは音素集合の似た言語同士では非効率的であり，似た音素が異なる音素としてモデリングされることで単言語でしか学習されない音素量が増加する．ここでは，ほぼ音素集合が重なる言語である，日本語と韓国語を取り上げ，対象言語を韓国語とし，日本語の音素を韓国語の音素に変換することで，以下の2,3,4 節の 3 ステップで同時学習を行う．

2 音素の変換

日本語の方が母音の数が少なく，日本語の音素セットは，韓国語のそれのほぼサブセットとなりうることから，日本語の音素を以下の変換ルールにより，韓国語の音素に変換した．韓国語の音素は後述の zeroth_korean で用いている体系とした．母音/a/ /i/ /u/ /e/ /o/は，変換せずそのままとした．ただし「す」「つ」「ず」の/u/は/eu/とした．「ん (/N/)」に関しては，直後の音素に応じて/m/ /ng/ /n/とした．「や」「ゆ」「よ」および拗音の「ゃ」「ゅ」「ょ」は/ya/ /yu/ /yo/で，その他/k/→/kh/，/sh/→/s/，/r/→/l/，/z/→/j/，/f/→/h/，/ts/→/ch/とした．促音に関しては無音のパッチム/d2/で代用した．長音は韓国語には存在しないので，無視して一つの母音としてモデル化する場合と，同じ母音を重ねて二重母音としてモデル化した場合を比較した．

Korean Acoustic Model Training by Using Japanese Speech Data
† Yuuki Tachioka, Denso IT laboratory

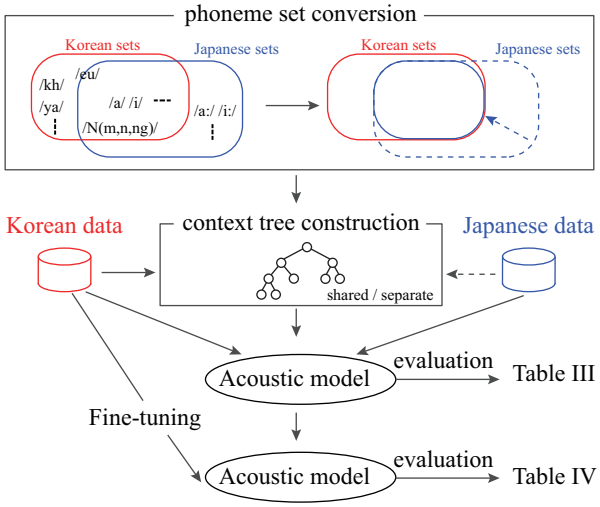

図 1 Simultaneous training procedure on both Korean (target) and Japanese (auxiliary) datasets. First step (Sec. 2) converts Japanese phonemes into Korean phonemes and the second step (Sec. 3) constructs shared or separate context trees. Afterwards, acoustic models are simultaneously trained on Korean and Japanese. Finally, acoustic models are fine-tuned only on the Korean dataset (Sec. 4).

3 コンテキスト木の構築

言語によって音素間の識別の重要性が異なる．例えば日本語では「ん (/N/)」は，/m/,/ng/,/n/の 3 通りに発音されるものの意味的な差異はないが，韓国語ではそれらは明確に区別され意味的な差異も生じる．そのため，コンテキスト木は対象言語のみで構築したほうが良い可能性がある．そこでコンテキスト木を両方の学習データから構築する場合と，韓国語の学習データだけから構築する場合を比較する．

4 対象言語でのファインチューニング

複数言語で同時に学習した汎用モデルのまま使うか，対象言語でファインチューニングを施すかどうかを比較する．ファインチューニングにより，両言語で重なり合う音素は学習データを増やすことができる一方，対象言語に特有の音素の識別性能の低下を避けられる．

表 1　Details of zeroth_korean corpus and CSJ.

	data length	# of speakers
training data		
zeroth_korean	51.6 hours	105 speakers
CSJ	239 hours	986 speakers
evaluation data		
zeroth_korean	1.2 hour	10 speakers

5　実験 (zeroth_korean データセット +CSJ)

5.1　実験設定

実験は zeroth_korean データセット[*1]を用いた．対象言語 (韓国語) の学習データは 51.6 時間である．補助言語の学習データとして，これより多い日本語話し言葉コーパス (CSJ)(239 時間) のデータを利用した．表 1 にデータの詳細を示す．Kaldi toolkit[*2]付属のスクリプトにより nnet3 で実験を行った．話者適応後の特徴量に対して DNN-HMM のハイブリットモデルにより認識を行う．言語モデルは規模の小さい tri-gram (tgs)，大きい tri-gram (tgl)，four-gram (fgl) の 3 種類で実験した．

5.2　結果と考察

表 2 に結果を示す．1 段目がベースラインである．これに対して 2 段目は韓日同時学習を行ったものであるが，コンテキスト木の構築方法，長音のモデル化法に関わらず，ベースラインよりも性能が低下している．学習話者が増えることにより，話者適応の性能が向上し認識性能も向上することが期待されたが，この効果は見られなかった．

3 段目は韓日同時学習ののちに韓国語だけでファインチューニングを行ったもので，ベースラインよりも性能が向上した．これより，汎用モデル学習後に，対象言語でファインチューニングを行う必要があることが分かった．この際，長音のモデル化法は，特に tgs 以外の言語モデルの場合，性能にあまり影響しなかった．コンテキスト木に関しては韓国語だけから構築したものの方が性能がよく，対象言語の認識を行うには上述の通り言語によって識別に重要な音素が異なるため，同時学習の際にも対象言語の音素体系を維持することが重要であることが分かった．

[*1] https://github.com/goodatlas/zeroth
[*2] https://kaldi-asr.org

表 2　WER[%] on the zeroth_korean corpus.

	tri-gram		four-gram
長音	small (tgs)	large (tgl)	large (fgl)
– zeroth_korean のみから学習 –			
	17.25	10.60	10.15
– 韓日同時学習 –			
コンテキスト木を韓国語だけから構築			
二重母音	20.08	11.77	11.43
無視	18.99	11.46	11.16
コンテキスト木を韓日で構築			
二重母音	20.96	11.76	11.27
無視	19.45	11.03	10.48
– 韓国語でファインチューニング –			
コンテキスト木を韓国語だけから構築			
二重母音	16.11	**9.88**	**9.42**
無視	**15.94**	9.90	9.48
コンテキスト木を韓日で構築			
二重母音	16.12	10.09	9.73
無視	16.36	10.00	9.83

6　まとめ

韓国語の音声認識性能向上のため，日本語の音素セットをルールにより韓国語の音素に変換して同時学習した．長音は無視しても二重母音としてもよく，コンテキスト木を韓国語だけから構築し，ファインチューニングすることで，韓国語のデータのみを用いた場合に比べて 0.7–1.3 ポイントの WER 改善が得られた．

参考文献

[1] S. Hara and H. Nishizaki, “Acoustic modeling with a shared phoneme set for multilingual speech recognition without code-switching,” Proc. APSIPA, pp.1–4, 2017.

[2] S. Tong, P.N. Garner, and H. Bourlard, “Multilingual training and cross-lingual adaptation on CTC-based acoustic model,” Speech Communication, vol.104, 2017.

[3] S. Watanabe, T. Hori, and J.R. Hershey, “Language independent end-to-end architecture for joint language and speech recognition,” Proc. ASRU, pp.265–271, 2017.