Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 167 – 173

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016
9-12 May 2016, Yogyakarta, Indonesia

Towards Robust Indonesian Speech Recognition with Spontaneous-Speech Adapted Acoustic Models

Devin Hoesen*, Cil Hardianto Satriawan, Dessi Puji Lestari, Masayu Leylia Khodra

Institut Teknologi Bandung, Jl. Ganeca No. 10, Bandung 40115, Indonesia

Abstract

This paper presents our work in building an Indonesian speech recognizer to handle both spontaneous and dictated speech. The recognizer is based on the Gaussian Mixture and Hidden Markov Models (GMM-HMM). The model is first trained on 73 hours of dictated speech and 43.5 minutes of spontaneous speech. The dictated speech is read from prepared transcripts by a diverse group of 244 Indonesian speakers. The spontaneous speech is manually labelled from recordings of an Indonesian parliamentary meeting, and is interspersed with noises and fillers. The resulting triphone model is then adapted only to the spontaneous speech using the Maximum A-posteriori Probability (MAP) method. We evaluate the adapted model using separate dictated and spontaneous evaluation sets. The dictated set consists of speech from 20 speakers totaling 14.5 hours. The spontaneous set is derived from the recording of a regional government meeting, consisting of 1085 utterances totaling 48.5 minutes. Evaluation of a MAP-adapted spontaneous set yields a 2.60% absolute increase in Word Accuracy Rate (WAR) over the un-adapted model, outperforming MMI adaptation. Conversely, MMI adaption of the dictated set outperforms the MAP adaptation by achieving an absolute increase of 1.48% in WAR over the un-adapted model. We also demonstrate that fMLLR speaker adaptation is unsuitable for our task due to limited adaptation data.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: **GMM-HMM; MAP adaptation; MMI adaptation; Spontaneous speech**

* Corresponding author.
E-mail address: 23514103@std.stei.itb.ac.id

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.045

1. Introduction

Spontaneous speech recognition has long been the most important and challenging task in speech recognition[12]. Dictated speech, such as in news broadcasts, can be recognized by GMM-HMM-based systems with an accuracy exceeding 90%. However, the same systems struggle to perform when employed to recognize spontaneous speech.

A number of explanations have been put forward to account for this performance degradation; it has been suggested that spontaneous speech differs from dictated speech linguistically and spectrally[7]. Linguistically, spontaneous speech is characterized by the inclusion of filled pauses, word or phrase repetition, interjections, unknown or mispronounced words, ellipsis (omission of pronouns and/or relatives), and ungrammatical sentences or unusual word order.

Current Indonesian speech research is focused mainly on improving recognition accuracy for dictated speech[3,5,13,15], primarily due to the scarcity of Indonesian speech resources, both dictated and spontaneous. It generally requires less effort to build a phonetically rich and balanced dictated speech corpus from a prepared set of transcripts than manually annotating spontaneous speech from recordings we have no control over.

Considering the difficulties of building a phonetically rich and balanced spontaneous speech corpus, an adaptation approach is used; instead of attempting to build a large, representative spontaneous speech corpus, a large amount of clean dictated speech data is collected to account for phonetic variation, is acoustically modelled, and then adapted to a small amount of spontaneous speech gathered in the specific task and acoustic environment we wish to deploy for.

Adaptation of acoustic models is employed to increase the recognition accuracy of a generic acoustic model. Adaptation can be employed to make the model more suitable to the speaker, the environment, or the task. Common adaptation techniques in the existing literature for GMM-HMM-based acoustic modelling include:

- **Maximum A-posteriori Probability** (MAP) adaptation provides a way to incorporate prior information to the training process and is useful if training/adaption data is sparse[6].
- **Maximum Mutual Information** (MMI) adaptation is a discriminative training/adaptation method for an HMM. To estimate the new parameter of an HMM, one must maximize the numerator term and minimize the denominator term in the objective function[11].
- **Feature-space Maximum Likelihood Linear Regression** (fMLLR) is the most commonly used speaker adaptation technique. In fMLLR, the means and variances of the HMM are transformed using a linear transformation. Since it is applied in feature-space only, it is a type of constrained model-space transformation[8].

For the task adaptation, MAP and MMI are claimed to be able to reduce the word error rate significantly[9]. In this paper, we employ the MAP and MMI techniques to adapt the acoustic model to a different task and environment.

2. Speech corpus description

The training set consists of a dictated and a spontaneous component. The evaluation set similarly consists of a dictated set and a spontaneous set. Each corpus is described in detail in this section.

2.1. Dictated speech training set

The dictated speech training set, READ-TR, was recorded in two phases. The first phase collected speech recorded by a diverse range of Indonesian speakers from 1 of 25 prepared transcripts, each consisting of 250 sentences. The second phase consists of speech read from one of nine prepared transcriptions, each consisting of 300 sentences.

Table 1. Summary of READ-TR.

Phase	#Speaker		#Sentences	Duration
	Male	Female		
First	99	99	49,493	48h18m36s
Second	25	21	13,798	24h37m14s
Total	124	120	63,291	72h55m50s

Utterances were recorded in 16-bit single-channel WAV at a sampling rate of 48 kHz in an acoustically clean studio environment with a signal-to-noise ratio (SNR) exceeding 30 dB. Data cleaning and preparation tasks include the manual checking of recorded utterances against their transcripts, the removal of redundant and corrupted files, and resampling of speech files into 16 kHz WAV. In total, the READ-TR corpus consists of 72 hours, 55 minutes, and 50 seconds of speech distributed over 63,291 files. The details of the READ-TR set are summarized in Table 1.

2.2. Spontaneous speech training set

The spontaneous speech training set, SPONTAN-TR, was recorded from an Indonesian parliamentary meeting discussing the preservation of regional languages. This particular session was chosen for its relatively clean speech segments, with mostly clear sentence endings, distinguishable speaker changes, and minimum speech overlap between speakers.

The recording was automatically segmented by short-time energy level. The resulting segments are interspersed with various intermittent noises such as breathing (rumbling), coughing, whistling (sibilance), and fillers. The recording was then manually annotated (labelled) to obtain the final training transcript. All recordings are in 16-bit single-channel WAV resampled to 16 kHz.

2.3. Dictated speech evaluation set

Standardized evaluation tasks do not yet exist for the Indonesian language. The dictated speech evaluation set, READ-EV, was obtained from the Lestari's research[5] with some modifications. A phonetically rich and balanced transcript consisting of 343 sentences was read by all speakers. As with the READ-TR set, this set was recorded in an acoustically clean studio environment to reduce environmental background noise. As speakers were recorded in a single pass, an additional segmentation step was required to obtain the individual segments.

Additional preparation steps were performed; speech files were manually checked against their respective transcripts and the transcripts corrected as necessary, and redundant and unintelligible speech files were eliminated. In total, the READ-EV corpus contains 6,682 speech segments for a total duration of 14 hours, 31 minutes, and 45 seconds.

2.4. Spontaneous speech evaluation set

The spontaneous speech evaluation set, SPONTAN-EV, is derived from recordings of regional government meetings. As with the SPONTAN-TR set, speech is segmented automatically based on short-time energy level, and is interspersed with various noises and fillers. The speech segments obtained from this process were then manually annotated to produce the final evaluation transcript. The final SPONTAN-EV set consists of speech from 23 male and 5 female Indonesian speakers for a total duration of 48 minutes and 38 seconds over 1,085 speech segments.

2.5. Summary

The details of each corpus are summarized in Table 2. The UNION-TR corpus denotes the joint READ-TR SPONTAN-TR corpus utilized in the triphone training process.

Table 2. Summary of the speech corpora.

Corpus	#Speakers	#Sentences	Duration
READ-TR	244	63,291	72h55m50s
SPONTAN-TR	11	1,058	43m35s
UNION-TR	255	64,349	73h39m25s
READ-EV	20	6,682	14h31m45s
SPONTAN-EV	28	1,085	48m38s

3. Experiment setup

A GMM-HMM speech recognition system typically utilizes a language model and lexicon to generate word hypotheses. This poses a significant challenge in the recognition of under-resourced languages such as Indonesian, for which a standard phonetic dictionary does not exist. In the following sections, we briefly discuss our efforts in building an Indonesian large vocabulary lexicon and language model, and describe in greater detail the final configuration used for building the acoustic models and their evaluation.

3.1. Building the dictionary

Dictionary terms were collected from the various training corpora transcripts (READ-TR phases one and two, SPONTAN-TR), the official Indonesian language dictionary (Indonesian: KBBI), and a large collection of news articles from reputable Indonesian websites. A large amount of preparation and cleaning was necessary, involving the addition of alternative pronunciations as heard in the various speech corpora, the standardization of term pronunciations, and the fixing of numerous spelling and tokenization errors in the crawled news articles. 182,309 unique terms were obtained in this manner.

Table 3. List of all phonemes utilized in the HMM training.

Our symbol	IPA symbol	Our symbol	IPA symbol	Our symbol	IPA symbol
a	/a/	i	/i/	q	/q/
ai	/aɪ̯/	j	/dʒ/	r	/r/
au	/aʊ̯/	k	/k/	s	/s/
b	/b/	kh	/x/	sy	/ʃ/
c	/tʃ/	l	/l/	t	/t/
d	/d/	m	/m/	u	/u/
e	/e/	n	/n/	w	/w/
@	/ə/	ng	/ŋ/	y	/y/
f	/f/	ny	/ɲ/	z	/z/
	/v/	o	/o/	SIL	
g	/g/	oi	/oɪ̯/	SPN	
h	/h/	p	/p/	NSN	

The pronunciation dictionary was compiled using the phoneme set shown in Table 3 and is a modification of the phoneme set described by Soderberg and Olson[2]. The /q/ phoneme is supplied to accommodate the pronunciation of certain Arabic loanwords containing the letter [q]. Additional phonemes are supplied to accommodate silence and noises. The /ʔ/ phoneme is removed as its Indonesian pronunciation is unclear. Other phonemes are supplied to accommodate silence and noises; SIL represents silence, SPN spoken noise, and NSN non-spoken noise. In total there are 32 non-silence phonemes.

3.2. Building the language model

The 3-gram language model (LM) used for acoustic model evaluation was built using the SRILM Language modelling Toolkit[1] on training data from the 2003 Tala text corpus[5]. The dictionary obtained in section 3.1 was used to obtain the list of words. Based on the results of Chen and Goodman's research[14], Kneser-Ney smoothing and the interpolation of higher-order with lower-order probability estimates was used to train the dictated speech 3-gram language model READ-LM.

In order to incorporate spontaneous speech traits, we further interpolated READ-LM with an LM trained on the SPONTAN-TR transcript to produce the spontaneous speech SPONTAN-LM language model. The optimum mixing weight of the two language models was assessed using SRILM for both the dictated and spontaneous scenario

(READ-EV's and SPONTAN-EV's transcript are utilized as the evaluation set for dictated and spontaneous scenario, respectively). To balance the perplexity for both scenarios, the optimum weights for each scenario, as shown in Table 4, were averaged and the result used to produce the final INTER-LM model. The INTER-LM model is used to evaluate both scenarios. Table 5 shows the statistics of the training and evaluation text corpus and Table 6 shows the summary of the computed perplexities over the evaluation text set.

Table 4. Best mixing weight for both scenarios.

Scenario	READ-LM	SPONTAN-LM	Total
Dictated	0.994006	0.005994	1
Spontaneous	0.542802	0.457198	1
Mean	0.768404	0.231596	1

Table 5. Statistics of the training and evaluation text.

Field	Tala 2003	SPONTAN-TR	READ-EV	SPONTAN-EV
#Unique Sentences	613,054	1,052	343	1,085
#Words	10,250,367	5,265	4,016	5,684
OOV Rate			4.18%	0.58%

Table 6. Perplexities for both scenarios.

LM	Dictated ppl	Spontaneous ppl
READ-LM	219.76	887.35
SPONTAN-LM	18,038.90	1,325.39
INTER-LM	240.54	469.30

3.3. Building the acoustic model

The Kaldi toolkit for speech recognition[4] is employed to train and later decode our GMM-HMM acoustic model. The toolkit is used to conduct feature extraction, triphone training, MAP adaptation, MMI adaptation, and fMLLR training on both the resulting MAP and MMI models, followed by the evaluation of each resulting acoustic model.

For every corpus set described in Section III, the standard 39 Mel Frequency Cepstral Coefficients are extracted. Cepstral Mean and Variance Normalization (CMVN) is then performed for each speaker. CMVN is known to normalize speech such that similar segmental parameter statistics are generated in all noise conditions[10].

The monophone system is first trained on the UNION-TR set. After alignment, the number of leaves and GMM mixtures for the triphone system is tuned to obtain optimum recognition rates, with Kaldi automatically allocating the number of mixtures per leaf. Subsequently, the number of leaves is varied, keeping the leaf to total-mixtures ratio constant. The leaf to total-mixtures ratio is then calculated for the optimum number of leaves in both scenarios. Table 7 summarizes our findings in this respect.

Based on the experiment, we choose the optimum values for the spontaneous over the read scenario. We train a triphone system with 3000 leaves and 96000 mixtures on the UNION-TR set. After alignment of the triphone model with the training set, we adapt the triphone to the SPONTAN-TR set using both MAP and 4-iteration MMI adaptation. We align the resulting MAP and MMI models with the SPONTAN-TR set and train an fMLLR system from each model, also with 3000 leaves and 96000 mixtures.

To decode with the model, we first convert the ARPA-formatted INTER-LM generated by SRILM to the Kaldi WFST format. The decoding is then done using the triphone, MAP, MMI, MAP-fMLLR, and MMI-fMLLR models on both the READ-EV and SPONTAN-EV sets.

4. Experiment results and analysis

4.1. Acoustic model evaluation result

The results of our experiment are shown in Table 7 and expressed in Word Accuracy Rate (WAR). WAR is calculated using Equation (1) where S is the number of substitutions, I is the number of insertions, D is the number of deletions, and N is the total number of words in the test set.

$$WAR = \left(1 - \frac{S + I + D}{N}\right) \times 100\% \tag{1}$$

4.2. Result analysis and discussion

From Table 7, the MAP-adapted triphone model is most accurate in the spontaneous scenario, while the MMI-adapted model performs better in the dictation scenario. However, when both models are adapted with speaker-dependent fMLLR, dictation performance degrades significantly. This is a result of the differing spectral characteristics of the speakers in the spontaneous scenario as opposed to the dictated scenario. These results demonstrate the suitability of MAP and MMI adaptation for small amounts of adaptation data and task and environment adaptation. On the other hand, fMLLR adaptation on insufficient adaptation data leads to suboptimal performance.

Table 7. Experiment result in % WAR.

Acoustic Model	SPONTAN-EV	READ-EV
Triphone	54.26	81.53
Triphone + MAP	**56.86**	82.89
Triphone + MMI	53.80	**83.01**
Triphone + MAP + fMLLR (SI)	41.19	21.46
Triphone + MAP + fMLLR (SD)	56.79	56.68
Triphone + MMI + fMLLR (SI)	42.19	19.88
Triphone + MMI + fMLLR (SD)	56.74	56.68

Note: SI = speaker-independent model
SD = speaker-dependent model

Closer inspection of the generated word hypotheses reveals that our adapted models are unable to recognize most filler words, such as <*euh*>, and all non-spoken noises such as breathing sounds. This is due to the disproportionately small amount of filler and non-spoken noise in the SPONTAN-TR training set. Furthermore, all occurrences of filler and non-spoken noise are grouped over-broadly into the spoken (SPN) and non-spoken (NSN) phonemes, respectively. Hence, the representation of filler and non-spoken noises in the acoustic model is probably over-generalized.

As seen in Table 6, the language model perplexities for both scenarios are high, with spontaneous speech more significantly affected than dictated speech. This is presumably due to incompatibilities between the domains of the training and evaluation sets; utterances in the training set are taken mostly from news articles whereas utterances in the spontaneous evaluation set are annotated from meetings. In particular, filler words are severely under-represented in the language model.

A large number of errors are caused by improper word segmentation and concatenation in the transcripts and text corpus, an example being the segmentation of the word *kedua* ‘second’ into *ke dua* ‘to two (sic)’. An example of improper word concatenation is the combination of the words *di sini* ‘here’ into *disini*. This is mainly due to the shared spelling of certain prefixes and prepositions, such as the prefix *ke-* and preposition *ke* ‘to’ in the example above.

There are also several occurrences of semantically identical words with alternative, but very similar, spellings. This problem is usually exhibited by named entities, such as *Mulyoharjo* and *Mulyohardjo*. This problem is aggravated by the modernization of the spelling of a number of phonemes, specifically of [dj] to [j] /dʒ/, [j] to [y] /j/,

[oe] to [u] /u/, and [tj] to [c] /tʃ/. Numbers can also fall into this category. Arabic number(s) can be decoded into Roman number(s) and vice versa.

A large number of recognition errors are also caused by similar sounding word sequences, such as *siap take off* 'ready to take off' recognized as *siap ceko* 'ready to Czech (sic)'. This can be remedied by training the LM on suitable sentences.

Lastly, the presence of un-adapted foreign and loan words, mainly from English and Arabic, and their misspelled variations in the Indonesian language, contributes significantly to the errors explained above. The Arabic greeting *Wassalamualaikum* and its misspelled variation *wasalamualaikum* is the most egregious example. All of the above are indications of an insufficiently clean training set, both for acoustic modelling and language modelling.

5. Conclusion and future works

A general-purpose acoustic model is built to recognize both dictated and spontaneous speech. MAP adaptation results in a 2.60% and 1.36% absolute increase in word accuracy rates (WAR) over the un-adapted model for spontaneous and dictated speech, respectively. For the dictated scenario, MMI adaptation yields an absolute increase of 1.48% in WAR. On the other hand, fMLLR adaptation on top of both models degrades performance significantly in both scenarios. Thus, we conclude that MAP and MMI adaptation is suitable when adaptation data is scarce, as is the case with under-resourced languages.

Future work includes determining whether performance improvements due to adaptation transfer over to DNN systems. More effective or appropriate methods for language model cleaning and interpolation are also required to reduce the large perplexities in both scenarios.

Acknowledgements

This research is partially supported by "Riset Unggulan Perguruan Tinggi Kemenristekdikti" (University Distinguished Research) and is part of a larger project titled "Pengembangan Perangkat Lunak Perisalah Rapat dengan Memanfaatkan Teknologi Pengenal Ucapan, Mesin Translasi, dan Peringkas Otomatis" (Development of Meeting Speech Transcriber Using Speech Recognition, Machine Translation, and Automatic Summarizer Technology). We also would like to thank PT INTI, Bandung, Indonesia for the utilization of the speech corpus presented in this research.

References

1. Stolcke A, SRILM – An extensible language modeling toolkit. *Proc. ICLSP* 2000.
2. Soderberg C.D, Olson K.S, Indonesian. *Journal of the IPA* 2008; **38**:2-209.
3. Hoesen D. et al, A DNN-based ASR system for the Indonesian language. *Proceedings ASJ Autumn Meeting* **2015**; pp. 5.
4. Povey D. et al, The Kaldi speech recognition toolkit. *IEEE Workshop on ASRU* **2011**.
5. Lestari D.P, Iwano K, Furui S, A large vocabulary continuous speech recognition system for Indonesian language. *15th Indonesian Scientific Conference in Japan Proceedings* **2006**.
6. Gauvain J.L, Lee C.H, Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE TSAP* **1994**; 2:2-291.
7. Nakamura M, Iwano K, Furui S, Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *ISCA CSL* **2008**; 22:2-171.
8. Gales M.J.F, Maximum likelihood linear transformations for HMM-based speech recognition. *ISCA CSL* **1998**; 12:2-75.
9. Gales M.J.F et al, Porting: Switchboard to the Voicemail Task. *IEEE ICASSP* **2003**; 1, pp. I-536.
10. Viikki O, Laurila K, Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication* **1998**; 25-133.
11. Woodland P.C, Povey D, Large scale discriminative training for speech recognition. *ISCA CSL* **2002**; 16:1-25.
12. Furui S, Recent advances in spontaneous speech recognition and understanding. *ISCA & IEEE Workshop on SSPR* **2003**.
13. Sakti S et al, Development of Indonesian large vocabulary continuous speech recognition system within A-STAR project. *Proceedings of the Workshop on TCAST* **2008;** pp. 19.
14. Chen S.F, Goodman J, An empirical study of smoothing techniques for language modelling. *ISCA CSL* **1999**; 13:4-359.
15. Ferdiansyah V, Purawarianti A, Indonesian automatic speech recognition system using English-based acoustic model. *AJSP* **2012;** 2:4-60.
16. Ward W, Understanding spontaneous speech. *ACL Speech and Natural Language: Proceedings of a Workshop* **1989**; pp. 137.