Building Open Javanese and Sundanese Corpora for Multilingual Text-to-Speech

Jaka Aris Eko Wibawa[2], Supheakmungkol Sarin[1], Chenfang Li[3], Knot Pipatsrisawat[1], Keshan Sodimana, Oddur Kjartansson[1], Alexander Gutkin[1], Martin Jansche[1], Linne Ha[1]
[1]Google,[2]Google (on contract from Pactera Technologies NA Inc),[3]Google (on contract from Optimum Solutions Pte Ltd)
{jwibawa,mungkol,chenfangl,thammaknot,ksodimana,oddur,agutkin,mjansche,linne}@google.com

Abstract

We present multi-speaker text-to-speech corpora for Javanese and Sundanese, the second and third largest languages of Indonesia spoken by well over a hundred million people. The key objectives were to collect high-quality data in an affordable way and to share the data publicly with the speech community. To achieve this, we collaborated with two local universities in Java and streamlined our recording and crowdsourcing processes to produce corpora consisting of 5,800 (Javanese) and 4,200 (Sundanese) mixed-gender recordings. We used these corpora to build several configurations of multi-speaker neural network-based text-to-speech systems for Javanese and Sundanese. Subjective evaluations performed on these configurations demonstrate that multilingual configurations for which Javanese and Sundanese are trained jointly with a larger corpus of Standard Indonesian significantly outperform the systems constructed from a single language. We hope that sharing these corpora publicly and presenting our multilingual approach to text-to-speech will help the community to scale up text-to-speech technologies to other lesser resourced languages of Indonesia.
Keywords: low-resource languages, corpora, multilingual, text-to-speech

1. Introduction

One of the important trends in modern speech and language technology is the increasing focus on scaling up the current state of the art to the large number of language communities in the world. Progress in this direction is difficult, because, more often than not, the languages in the long tail of the distribution of the majority of the world's languages lack adequate linguistic resources required for supporting speech and natural language research (Besacier et al., 2014; O'Horan et al., 2016).

A prime example of such a challenge are the languages of Indonesia. Lewis et al. (2015) identify 707 distinct living languages (not dialects) that are spoken throughout the archipelago by over 255 million people. About twenty of these languages are spoken by over a million people. Standard Indonesian is the official national language of Indonesia and is spoken natively or as a second language by more than 200 million people (Paauw, 2009). Javanese with roughly 90 million native speakers and Sundanese with approximately 40 million native speakers constitute the two largest regional languages of Indonesia. Unlike Indonesian, which received a lot of attention over the years, e.g. (Sakti et al., 2008; Manurung et al., 2010; Koto, 2016), both Javanese and Sundanese are currently under-resourced due to the lack of openly available high-quality corpora.

We address this problem for Javanese and Sundanese by building crowdsourced text-to-speech corpora for these languages. This work is part of an ongoing effort by Google to build multi-speaker corpora for low-resource languages in an affordable way and to share the resulting data publicly. Building on our initial work on Bengali (Gutkin et al., 2016), Sinhala and other languages, we aimed to design the process from the ground up, starting with script building and recording equipment selection to logistics and operations. Similar to other languages, we are releasing the resulting corpora under a liberal license on OpenSLR[1].

We collaborated with two universities in Indonesia to conduct data collections locally. For Javanese we worked with Universitas Gadjah Mada (UGM), Faculty of Cultural Science in Yogyakarta. For Sundanese we worked with Universitas Pendidikan Indonesia (UPI), Faculty of Language and Literature Education in Bandung. The universities assisted us with finding volunteers to help manage the data collection, as well as with adequate recording environments. Together with several researchers from Reykjavík University in Iceland we also used this opportunity to collect open-source Javanese and Sundanese[2] data for automatic speech recognition (ASR) (Guðnason et al., 2017).

We used the collected multi-speaker data to build statistical parametric speech synthesis (SPSS) systems for Javanese and Sundanese. While it is possible to use single-language corpora to bootstrap text-to-speech for individual languages, we also experimented with an alternative approach that capitalizes on language similarities between Javanese and Sundanese on the one hand, and Standard Indonesian on the other. We hypothesize that constructing a jointly trained multilingual system may result in significant improvements over the systems constructed using a "classical", single-language approach.

2. Overview of Javanese and Sundanese

Since we are building speech corpora, a key consideration is the design of the phoneme inventory of each language. In order to facilitate multi-lingual experiments, the phoneme inventories of different languages have to be bridged somehow. In the present case this was particularly straightforward, as the languages are related (they all belong to the Malayo-Polynesian subgroup of Austronesian) and their phoneme inventories overlap heavily.

[1] http://www.openslr.org/
[2] http://www.openslr.org/{35,36}/

Language		*Segments*
Indonesian	id	32 see below (23 consonants, 9 vowels)
Sundanese	su	33 Indonesian plus ɤ
Javanese	jv	35 Indonesian plus ʈ ɖ ɔ
Joint total		36 (25 consonants, 11 vowels)

Consonants

p b	t d	ʈ ɖ	tʃ dʒ	k g		ʔ
m	n		ɲ	ŋ		
	r					
f	s z		ʃ	x		h
w	l		j			

Vowels

i			u
e	ə	ɤ	o
a			ɔ
	ɑɪ	ɑʊ	ɔɪ

Table 1: Joint phoneme inventory of Indonesian, Javanese, and Sundanese in International Phonetic Alphabet notation

We started from a pre-existing phoneme inventory of Standard Indonesian, which consists of 32 segmental phonemes. Table 1 shows the basic Indonesian inventory in IPA notation, with the additional segments for Javanese and Sundanese highlighted. The overall joint phoneme inventory consists of 36 segments.

Sundanese has one additional vowel phoneme, which is not shared with Indonesian or Javanese. This is the back unrounded vowel /ɤ/, written as *eu* in the modern Sundanese orthography. It occurs in words like *seueur* /sɤ.ɤr/ (many) and *henteu* /hən.tɤ/ (not) and contrasts with schwa /ə/.

Javanese has three additional phonemes in our analysis, which are not shared with Indonesian or Sundanese. One is the open back vowel /ɔ/. The other two are, remarkably, the retroflex stops /ʈ/ and /ɖ/.

The retroflex stops occur, inter alia, in loanwords that ultimately trace back to languages of India, for example *garudha* /ga.ru.ɖɔ/ (Garuda, eagle; from Sanskrit *garuḍa* or *kutha* /ku.ʈɔ/ (town; compare Indonesian *kota*). Here we also see that orthographic final *a* is read as phonemic /ɔ/. By contrast in *gajah* /ga.dʒah/ (elephant) the last vowel does not round to /ɔ/, due to the presence of a final consonant. Phonemic /ɔ/ further arises as the reading of orthographic *o* in closed syllables, potentially spreading to preceding syllables: compare *koyo* /ko.jo/ with *koyok* /kɔ.jɔʔ/.

Our phoneme inventory for Javanese is largely identical to the one described by Ogloblin (2005).[3] Ogloblin lists an additional open-mid front vowel /ɛ/, which is not reflected in our inventory. We chose to treat this distinct sound as an allophone of /e/, but this choice is debatable.

Our joint inventory inherited several choices made for Indonesian, which we decided not to revisit. For example, the diphthongs are rare and could easily be analyzed as a combination of two vowels or as a vowel plus off-glide. The diphthong /ɔɪ/ is exceedingly rare and occurs mostly in loanwords. Notably all inventories include /ʃ/ and /x/ as distinct phonemes, as they have a robust presence in loanwords (and have distinct letters in e.g. the traditional orthography). Relatively recent orthographic reforms have made the modern Javanese and Sundanese orthographies highly regular and phonemically transparent. However, in conventional usage several ambiguities arise. Most notably, and shared with Indonesian, the distinction between /ə/ and /e/ is not always reliably indicated in the orthography. While /e/ can be written as *é*, this is not always done systematically, and often it appears as plain *e*. As a result, the orthography alone is insufficient to derive accurate pronunciations for a text-to-speech system. We therefore asked native speakers to transcribe substantial pronunciation dictionaries for Javanese (54,000 words) and Sundanese (42,000 words).

3. Data Collection

3.1. Recording Script

To build the script efficiently, we asked native speakers to list some of the important named entities (e.g., local place names), time expressions (e.g., months of the year), numbers (e.g., all numbers smaller than 100) and so on.

For Javanese, we asked the native speakers to create sentences that include these elements. The sentences should be easy to read, rich enough to include orthographic variants, not offensive, and span five to twenty words in length. For Sundanese, we scaled the process further by starting from *templates* constructed from elements in the the above lists. For example, "[global celebrity name] goes to [global place name city] with [local celebrity name] during [time expression season]" is one such template where the fillers are shown in square brackets. We then generated sentences from these templates, which were reviewed by native speakers. Finally, we computed grapheme and phoneme coverage of the resulting sentences to make sure that we cover most sounds of the language.

3.2. Hardware and Recording Setup

The equipment used was an ASUS Zenbook UX305CA fanless laptop, a Neumann KM 184 microphone, Blue Icicle XLR-USB A/D converter and a portable acoustic vocal booth[4]. Parameters such as distance between speaker and microphone, height of the microphone and the angle at which it is pointed to the speaker were kept as constant as possible.

All audio was recorded using ChitChat, our in-house recording tool described in (Gutkin et al., 2016). ChitChat is a web-based recording software that allows audio data to be collected, managed and quality controlled. Each volunteer is presented with a series of sentences assigned to them for recording. The tool records at 48 kHz (16 bits per sample), detecting audio clipping to ensure quality, and ambient noise prior to recording each sentence, with a high noise level triggering an alert preventing further recording. Audio is initially stored on the client, and is uploaded asynchronously to a server when requested.

3.3. Crowdsourced Data Collection

The staff of the Javanese Literature Department of UGM put us in contact with the volunteers from that university who helped record samples. For Sundanese, we worked with UPI who helped us find volunteer speakers. A portion of the recordings was done at the student-run CompFest 2016 (an annual Computer Science exhibition event organized by students from the Faculty of Computer Science at

[3] http://phoible.org/inventories/view/1675. Their /y/ should be understood as IPA /j/.

[4] https://www.vocalboothtogo.com/

Gender	Collected Data		Cleaned Data	
	jv-ID	su-ID	jv-ID	su-ID
Female	3912 (20)	2475 (21)	2864 (19)	2401 (21)
Male	3918 (19)	2594 (21)	2958 (19)	1810 (21)

Table 2: Javanese and Sundanese database details.

Universitas Indonesia)[5]. Volunteers were between 18 and 35 years old.

We asked the speakers to read the script that we prepared. Each one spent an average of one hour contributing approximately 80 recordings. They were advised to read the script naturally. The Javanese recordings were done at UGM; half of the data was collected in a studio, the rest in a quiet room, inside a portable vocal booth. All the Sundanese recordings were collected in a portable vocal booth placed inside a quiet room. We performed quality control (QC) using ChitChat after the data was collected. Our objective was to make sure that the text matches the recordings and that the recordings are free of any artifacts: background noise, breathing, extraneous noises, etc.

Table 2 shows the details for the Javanese and Sundanese databases for female and male speakers after the data collection was complete (columns on the left) and after the recordings passed the QC (columns on the right). For each stage a number of recorded sentences along with the number of speakers (in parentheses) are shown. The total duration of the female multi-speaker datasets is 3.5 hours for Javanese and 3.2 hours for Sundanese. For the male multi-speaker datasets the total duration is 3.5 hours for Javanese and 2.2 hours for Sundanese.

4. Experiments

The experiments focus on building individual text-to-speech systems for Javanese and Sundanese languages from the multi-speaker data described in Section 3. The goals of the experiments are as follows: The first goal is to determine whether the individual corpora that we collected are good enough to construct a text-to-speech voice of adequate quality for the languages in question. The second goal is to see whether the quality of the systems can be improved by utilizing joint training, where both languages are trained together in a single model. In addition, we are interested in including a high-quality single-speaker Indonesian language database in this experiment. Our hypothesis is that the presence of a larger and professional-quality corpus from a related language should positively impact the training (Li and Zen, 2016; Gutkin, 2017).

4.1. Language Data and System Details

In this experiment we use the Javanese, Sundanese (both described in Section 3) and Indonesian datasets. The Indonesian dataset is the largest of the three, being approximately six times bigger than Javanese and Sundanese datasets. The original audio for all three corpora was recorded at 48 kHz. We selected the female multi-speaker datasets for both Javanese and Sundanese for the purpose of these experiments.

[5] http://www.compfest.web.id

Code	Description	Speaker	
		jv	su
id	Indonesian single speaker		
jv	Javanese single speaker	✓	
jv+id	Javanese with Indonesian	✓	
jv+su+id	Javanese with Sundanese and Indonesian	✓	✓
su	Sundanese single speaker		✓
su+id	Sundanese with Indonesian		✓
su+jv+id	Sundanese with Javanese and Indonesian	✓	✓

Table 3: Different acoustic model configurations.

We constructed several statistical parametric speech synthesis (SPSS) systems for Indonesian, Javanese, and Sundanese. Each system consists of a linguistic front-end, followed by an LSTM-RNN acoustic model, and finally a vocoder (Zen and Sak, 2015; Zen et al., 2016). The details of the LSTM-RNN model configuration (neural network training parameters, multilingual input and output features and so on) are similar to the ones described by Gutkin (2017). The main difference is that in this work we use fewer neural network parameters because we are dealing with a much smaller set of languages.

For each of the Javanese and Sundanese multi-speaker datasets we selected the "best" (according to our subjective analysis) sounding speaker to be used as the target speaker for biasing the acoustic models during run-time. The use of speaker features have been demonstrated to improve the quality of a multilingual multi-speaker synthesis (Li and Zen, 2016; Gutkin, 2017). Various LSTM-RNN configurations we built are described next.

4.2. System Configurations

Overall we constructed seven configurations corresponding to various combinations of languages shown in Table 3. We built a single-speaker Indonesian system (id) from an Indonesian corpus to see how well the Indonesian system fares on its own, as a reference point. Similarly, we used the Javanese and Sundanese multi-speaker corpora individually to construct single-language (but multi-speaker) systems jv and su, which serve as baselines in our experiments. In addition, we jointly trained Javanese with Indonesian (jv+id) and Sundanese with Indonesian (su+id) to produce the bilingual Javanese and Sundanese systems. Finally, we jointly trained a single combined Javanese, Sundanese and Indonesian acoustic model in order to produce a trilingual Javanese (jv+su+id) and Sundanese (su+jv+id) systems.

Each resulting system was evaluated using a subjective Mean Opinion Score (MOS) listening test. For each test we used 100 sentences not included in the training data for evaluation. Each rater was a native speaker of the language and was asked to evaluate a maximum of 100 stimuli. Each item was required to have at least three ratings. The raters used headphones. After listening to a stimulus, the raters were asked to rate the naturalness of the stimulus on a 5-point scale (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent). Each participant had one minute to rate each stimulus. The rater pool included eight raters for Javanese and seven raters for Sundanese. For each language, all configurations were evaluated in a single experiment. For Indonesian, which is the "bigger" reference language in our experiments, we

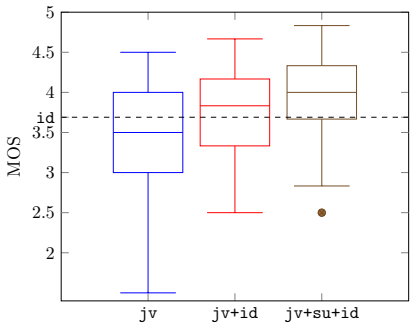

Figure 1: Mean Opinion Scores (MOS) box plot for three Javanese configurations.

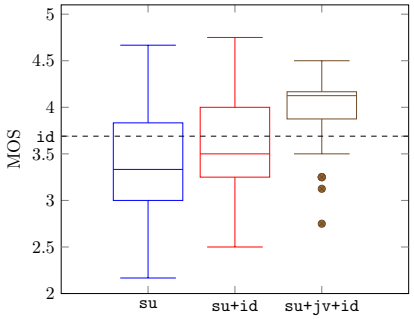

Figure 2: Mean Opinion Scores (MOS) box plot for three Sundanese configurations.

required over eight ratings per item and a larger rater pool.

4.3. Evaluation Results and Discussion

The distribution of the MOS scores for each of the three Javanese and three Sundanese systems described in the previous section are shown in Figures 1 and 2, respectively. The distributions are represented in box plot format, where, for each distribution the minimum, first quartile, median, third quartile, and maximum are displayed. Outlier values are shown as circles. The reference MOS score for Indonesian (`id`) is shown as a dashed line. The distributions are computed over all the available ratings for the 100 stimuli in each language.

Table 4 shows the mean opinion scores corresponding to the box plots for Javanese (Figure 1) and Sundanese (Figure 2). The mean opinion score for Indonesian is included for reference. Each mean opinion score is shown along with the corresponding 95% confidence interval (Recommendation ITU-T P.1401, 2012). The best scores are shown in bold.

As can be seen from the MOS box plots and Table 4, the baseline Javanese (`jv`) and Sundanese (`su`) systems bootstrapped solely from their respective multi-speaker corpora are inferior to the Indonesian system (`id`). In particular, the Sundanese system is significantly worse. This is not surprising because the Indonesian corpus is significantly larger, of studio quality, and consists of recordings from a single professional speaker.

These results, however, are improved when Javanese is jointly trained with Indonesian (`jv+id`) and Sundanese is jointly trained with Indonesian (`su+id`). These new configurations outperform both the baseline systems (`jv` and `su`) and their performance is tied with the Indonesian system (the differences with Indonesian are not statistically significant).

Configuration	Language	MOS
`id`	Indonesian	3.692±0.054
`jv`	Javanese	3.484±0.122
`jv+id`	Javanese	3.780±0.114
`jv+su+id`	Javanese	**3.998±0.103**
`su`	Sundanese	3.333±0.122
`su+id`	Sundanese	3.597±0.096
`su+jv+id`	Sundanese	**4.000±0.061**

Table 4: Subjective Mean Opinion Scores (MOS) (along with 95% confidence intervals) for languages synthesized with various acoustic model configurations. Best scores are shown in bold.

The best results are obtained by jointly training all the available languages. These Javanese (`jv+su+id`) and Sundanese (`su+jv+id`) systems outperform the systems built from two languages and also significantly improve upon the single-language multi-speaker baselines. It is interesting to note that despite using different speaker identity features and linguistic front-ends, the MOS scores for the two trilingual systems are virtually identical. We hypothesize that these additional improvements stem from both the availability of increased amounts of data from a related language (Indonesian) and from more data from a related language recorded in similar conditions (Javanese and Sundanese).

5. Conclusion and Future Work

We assembled open crowdsourced Javanese and Sundanese multi-speaker corpora collected for the purpose of building text-to-speech applications. Building on our previous work on Bengali (Gutkin et al., 2016) we further optimized the low-resource language data collection process to be more affordable.

We used the corpora to build several configurations of text-to-speech systems for Javanese and Sundanese. We demonstrated that the best results are obtained by constructing individual systems that share a multilingual acoustic model that is jointly trained on Javanese, Sundanese and a larger Indonesian dataset. This configuration significantly outperforms the baselines.

We hope that the process, the data, and the approaches described in this paper can be used to scale up our system to other low-resource languages of Indonesia, for example large languages such as Madurese (Davies, 2010) and Minangkabau (Adelaar, 1995).

One potentially very interesting venue for future research is to investigate different approaches to reducing the data collection costs by recording less data without sacrificing the synthesis quality.

6. Acknowledgments

The authors would like to thank Ruli Manurung, Maftuhah Ismail, Neneng Nurjanah, Mollyna Ezyando, Laila Rahmawati and Kharisma Ulinnuha for providing the linguistic expertise and for helping with the data collection. We also thank Universitas Gadjah Mada (UGM) and Universitas Pendidikan Indonesia (UPI) for their help with the recordings.

7. Bibliographical References

Adelaar, K. A. (1995). Minangkabau. In D. T. Tryon, editor, *Comparative Austronesian Dictionary 1*, pages 433–442. Mouton de Gruyter, Berlin / New York.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic Speech Recognition for Under-Resourced Languages: A Survey. *Speech Communication*, 56:85–100.

Davies, W. D. (2010). *A Grammar of Madurese*. Mouton Grammar Library. Mouton de Gruyter, August.

Guðnason, J., Pétursson, M., Kjaran, R., Klüpfel, S., and Nikulásdóttir, A. B. (2017). Building ASR corpora using Eyra. In *Proc. of Interspeech 2017*, pages 2173–2177, Sweden, August.

Gutkin, A., Ha, L., Jansche, M., Pipatsrisawat, K., and Sproat, R. (2016). TTS for Low Resource Languages: A Bangla Synthesizer. In *10th edition of the Language Resources and Evaluation Conference (LREC)*, pages 2005–2010, Portorož, Slovenia, May.

Gutkin, A. (2017). Uniform Multilingual Multi-Speaker Acoustic Model for Statistical Parametric Speech Synthesis of Low-Resourced Languages. In *Proc. of Interspeech 2017*, pages 2183–2187, Sweden, August.

Koto, F. (2016). A Publicly Available Indonesian corpora for Automatic Abstractive and Extractive Chat Summarization. In *10th edition of the Language Resources and Evaluation Conference (LREC)*, pages 801–805, Portorož, Slovenia, May.

Lewis, M. P., Simons, G. F., Fennig, C. D., et al. (2015). *Ethnologue: Languages of the world*, volume 20. Dallas, TX: SIL International, `http://www.ethnologue.com/`.

Li, B. and Zen, H. (2016). Multi-Language Multi-Speaker Acoustic Modeling for LSTM-RNN based Statistical Parametric Speech Synthesis. In *Proc. of Interspeech 2016*, pages 2468–2472, San Francisco, September.

Manurung, R., Distiawan, B., and Putra, D. D. (2010). Developing an Online Indonesian Corpora Repository. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*.

Ogloblin, A. K. (2005). Javanese. In Alexander Adelaar et al., editors, *The Austronesian Language of Asia and Madagascar*, pages 590–624. Routledge.

O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., and Korhonen, A. (2016). Survey on the Use of Typological Information in Natural Language Processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 1297–1308, Japan.

Paauw, S. (2009). One Land, One Nation, One Language: An Analysis of Indonesia's National Language Policy. *University of Rochester Working Papers in the Language Sciences*, 5(1):2–16.

Recommendation ITU-T P.1401. (2012). Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. *International Telecommunication Union*, July.

Sakti, S., Kelana, E., Riza, H., Sakai, S., Markov, K., and Nakamura, S. (2008). Development of indonesian large vocabulary continuous speech recognition system within A-STAR project. In *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*.

Zen, H. and Sak, H. (2015). Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, pages 4470–4474, Brisbane, Australia, April. IEEE.

Zen, H., Agiomyrgiannakis, Y., Egberts, N., Henderson, F., and Szczepaniak, P. (2016). Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices. In *Proc. of Interspeech*, San Francisco, September. ISCA.