Uniform Multilingual Multi-Speaker Acoustic Model for Statistical Parametric Speech Synthesis of Low-Resourced Languages

Alexander Gutkin

Google Inc., United Kingdom

agutkin@google.com

Abstract

Acquiring data for text-to-speech (TTS) systems is expensive. This typically requires large amounts of training data, which is not available for low-resourced languages. Sometimes small amounts of data can be collected, while often no data may be available at all. This paper presents an acoustic modeling approach utilizing long short-term memory (LSTM) recurrent neural networks (RNN) aimed at partially addressing the language data scarcity problem. Unlike speaker-adaptation systems that aim to preserve speaker similarity across languages, the salient feature of the proposed approach is that, once constructed, the resulting system does not need retraining to cope with the previously unseen languages. This is due to language and speaker-agnostic model topology and universal linguistic feature set. Experiments on twelve languages show that the system is able to produce intelligible and sometimes natural output when a language is unseen. We also show that, when small amounts of training data are available, pooling the data sometimes improves the overall intelligibility and naturalness. Finally, we show that sometimes having a multilingual system with no prior exposure to the language is better than building single-speaker system from small amounts of data for that language.

Index Terms: speech synthesis, low-resourced languages, long short-term memory, recurrent neural networks

1. Introduction

In recent years, statistical parametric speech synthesis has seen a shift of interest from Hidden Markov Models (HMMs) [1] to neural networks, starting with the work of Zen *et al.* [2], who demonstrated that feed-forward deep neural networks (DNNs) can achieve better naturalness than HMM systems. The performance of statistical parametric speech synthesis has further improved with the introduction of long short-term memory (LSTM)-based recurrent neural networks (RNNs) [3, 4] and, more recently, direct PCM (audio) generative models [5].

Despite recent advances in parametric speech synthesis, as well as wider availability of versatile tools for constructing speech synthesis systems, there still remain fundamental challenges. The primary challenge is the speech data acquisition, which can be a very labor-intensive and expensive process for constructing a high-quality system. This problem has been approached from several angles. Speaker adaptation techniques are suitable when some data from a prior collection is available and one needs to adapt this data to a new speaker using a small set of recordings [6, 7, 8]. However, this approach may not apply for under-resourced languages, when no source corpora are available. In this scenario, crowd-sourcing the data from multiple speakers and building an average voice is possible [9]. For the majority of the world's languages, in the long tail of the distribution [10], even these approaches may not be feasible, due to the lack of sufficient audio data, linguistic resources, or adequate infrastructure [11].

In this work we constrain the problem to a specific scenario where one is guaranteed to have some minimal linguistic representation of a particular (possibly under-resourced) language. This allows one to develop a (possibly very basic) linguistic front-end that outputs linguistic features which serve as an input to the acoustic model. The acoustic model is trained on multiple languages and may never observe the target language in its training data. This type of acoustic models, the multilingual multi-speaker (MLMS) models, were proposed in [12, 13, 14]. These approaches utilize a large input feature space consisting of "concatenated" language-dependent components, one for each language. In [13], there are several language-specific RNN output layers, whereas in [14] the multiple RNN output layers are speaker-specific and language-agnostic.

The approach taken in this paper is different: the goal is to investigate a *uniform* MLMS model which is both language- and speaker-agnostic, apart from a very limited fixed set of language- and speaker-identifying features. The input feature space takes any input from the linguistic front-end without encoding it as a sub-space of an input feature space, the internal layers have no language or speaker-specific structure and there is a single output layer. Controlled by the input features the output layer can generate acoustic parameters for any speaker, gender and language combination. Crucially, the input feature space does not need reconfiguration to support new languages. It is particularly interesting to investigate how well this model copes with "unseen" languages or languages for which very little amounts of training data is available.

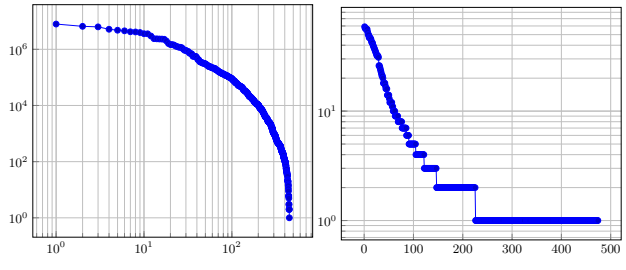
This paper is organized as follows: an overview of proposed multilingual model architecture is given in Section 2. Experiments and results are described in detail in Section 3. We conclude the paper in Section 4.

2. Multilingual Architecture

In this section we present the MLMS acoustic model that, together with the vocoder [15], forms a standalone back-end component in a multilingual text-to-speech system. The input to the back-end is a set of linguistic features. The output is streamed audio (linear PCM). The following properties set the proposed architecture apart from the ones recently reported in the literature [13, 14]: (1) A compact input feature space that does not need updating to support new languages; and (2) a simple network architecture similar to a single-speaker system. These properties are described in more detail next.

2.1. Linguistic features

Typically, the input to the acoustic model in statistical parametric speech synthesis consists of many diverse types of linguistic features. The features may include positional information (po-



(a) $\log(N_p)$ *vs.* $\log(r_p)$ (b) $\log(N_l)$ *vs.* r_p

Figure 1: *Rank r vs. frequency N plots for individual phonemes p: Global counts N_p in pronunciation dictionaries (a) and language counts N_l (b).*

sition of a phone in a syllable, number of syllables in a sentence, phrase finality), morphology (gender, number and case), syntax (marking intonational phrases), and so on.

More often than not, the set of linguistic features available for a low-resource language is limited by the lack of linguistic analysis components. For example, it is hard to find a reliable dependency tree parser for Burmese or a high-quality morphological analyzer for Belarusian. Nevertheless, it is often possible to build a minimal system without these features that is still "better than nothing".

2.1.1. Canonical phonological representation

The training set (described in detail in Section 3) contains diverse corpora for many languages and dialects following different phonological transcription conventions. In order to train the acoustic model on multilingual data, we transform each of the single-speaker phonemic configurations into a unified canonical IPA representation [16], similar to [12]. At present, this process is quite involved because it requires linguistic expertise for constructing the mappings for each of the individual languages. Additional difficulty presents itself when these mappings disagree due to differences between transcribers, diverging transcription conventions, or the lack of native speakers to guide the design. For example, /ɪ̃ʊ̃/ may not be the most accurate representation of a particular diphthong found in Nepali. Nevertheless, the canonical IPA representation yields a reasonably compact phonological description for many languages we deal with.

Figure 1a shows the logarithmic plot of rank (r_p) vs. frequency (N_p) distribution of canonical phonemes (p) obtained by the mapping (474 in total). These phonemes are encountered in pronunciation lexicons for 59 language/region pairs represented by phonological mapping and are ranked according to their frequency of occurrence. The degree of sharing of canonical phonemes p among 59 language/region pairs is shown in Figure 1b. Here the phonemes are ranked (r_p) according to the number of languages that share them (N_l). The plot shows reasonably high degree of sharing for approximately up to 70 phonemes shared by 10 languages or more. The coverage is poor for the long tail of about 250 phonemes that are only used in one language. This issue needs addressing by revisiting the phoneme inventories and choosing less sparse phoneme representation. However, we hypothesize that, due to partial decomposition of each phoneme into corresponding articulatory features (such as place of articulation) [17, 18], even the long-tail phonemes still contribute to the overall model.

2.1.2. Phylogenetic language features

We use language and region identifying features based on the BCP-47 standard [19] to model the similarity between various accents of the same language. This works well in practice for languages like English, where the language code enforces additional degree of similarity between American English (EN-US) and Australian English (EN-AU). This, however, is not sufficient for modeling the similarity between related languages. The language code for Slovak (SK), for example, tells us nothing about how it relates to Czech (CS).

In order to model one aspect of potential language similarity we employ a feature encoding of a traditional, if imperfect, phylogenetic language classification tree [20] that represents related clusters of languages down to a depth of four levels. Some languages in our representation – e.g. Hungarian – require three categorical features to encode their tree, while others – e.g. Marathi – require four categorical features.

2.2. LSTM-RNN acoustic model

We use LSTM-RNNs designed to model temporal sequences and long-term dependencies between them [21]. These types of models have been shown to work well in speech synthesis applications [3, 22, 23, 4]. Our architecture is very similar to one proposed by Zen *et al.* [4, 24]: unidirectional LSTM-RNNs for duration and acoustic parameter prediction are used in tandem in a streaming fashion. Given the linguistic features, the goal of the duration LSTM-RNN is to predict the duration (in frames) of the phoneme in question. This prediction, together with the linguistic features, is then given to the acoustic model which predicts smooth acoustic parameter trajectories. The smoothing of transitions between consecutive acoustic frames is achieved in the acoustic model by using recurrent units in the output layer.

Because we deal with significantly larger amount of training data and a more diverse set of linguistic features and recordings, the main difference between our model and the model in [4] is in the number of units in the rectified linear unit (ReLU) [25] and LSTM layers, as well as the number of recurrent units in the output layer of the acoustic model. The details of the duration and acoustic models are provided in the next section.

3. Experiments

The multilingual corpus used for training the acoustic models has over 800 hours of audio and consists of 37 distinct languages. These languages belong to the original group of 59 language/region pairs (described in Section 2) for which acoustic training data is available. Some languages, such as English, have different speaker datasets corresponding to different regional accents. For some accents (like EN-US) we have several speakers. Figure 2 shows on a logarithmic scale the relative distribution of languages within the training data in terms of number of training utterances. The distribution is heavily skewed towards "big" languages, and English in particular.

The corpus has both male and female speakers and is quite mixed: Most datasets are single-speaker, while others, like Bangladeshi Bengali and Icelandic, have recordings from multiple speakers [9]. The recording conditions vary: some speakers were recorded in anechoic chambers, while others in a regular recording studio setup or on university campuses. The F_0 range of the speaker varies from low F_0 males (Estonian) to high F_0 females (Korean). No speaker normalization was performed on the data.

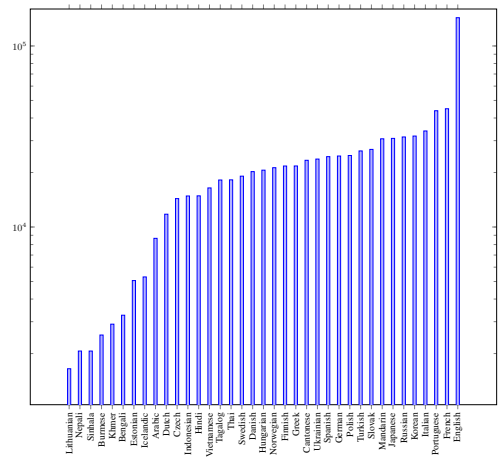


Figure 2: *Number of training utterances per language displayed on logarithmic scale.*

3.1. Methodology: System details

The speech data was downsampled to 22.05 kHz. Then mel-cepstral coefficients [26], logarithmic fundamental frequency ($\log F_0$) values (interpolated in the unvoiced regions), voiced/unvoiced decision (boolean value) [27], and 7-band aperiodicities were extracted every 5 ms, similar to [24]. These values form the output features for the acoustic LSTM-RNN and serve as input vocoder parameters [15]. The output features for the duration LSTM-RNN are phoneme durations (in seconds). The input features for both the duration and the acoustic LSTM-RNN are linguistic features. The acoustic model supports multi-frame inference [24] by predicting four frames at a time, hence the training data for the model is augmented by frame shifting up to four frames. Both the input and output features were normalized to zero mean and unit variance. At synthesis time, the acoustic parameters were synthesized using the Vocaine vocoding algorithm [15].

The architecture of the acoustic LSTM-RNN consists of 2×512 ReLU layers [25] followed by 3×512-cell LSTMP layers [28] with 256 recurrent projection units and a linear recurrent output layer [4]. The architecture of the duration LSTM-RNN consists of 1×512 ReLU layer followed by a single 512-cell LSTMP layer with a feed-forward output layer with linear activation. For both types of models the input and forget gates in each memory cell are coupled since distributions of gate activations for input and forget gates were previously reported as being correlated [29]. The duration LSTM-RNN was trained using an ϵ-contaminated Gaussian loss function [24], whereas for acoustic LSTM-RNN the L_2 loss function was used because we observed it to lead to better convergence rates.

3.2. Model configurations and evaluation

The experiments focus on two scenarios: In the first scenario, the model is trained on the corpus that excludes 12 languages (for six of which no acoustic data is available). The excluded group includes two Dravidian, two Indo-Aryan, one Romance, one Baltic, one North-Germanic, one Finno-Ugric and four Slavic languages. Each excluded language has some "relatives" remaining in the training data, apart from the Dravidian (Tamil and Telugu) and Baltic (Lithuanian) languages. At synthesis time, given the linguistic front-end features for each of the excluded languages, we synthesize the test sentences in the unseen languages using the above acoustic model (H, for "held-out") and verify the intelligibility of the output by subjective listening tests.

In the second scenario, we train an all-inclusive (I) model on all the data (37 languages). In this case we are primarily interested to find out whether training the language together with the others improves the overall synthesis quality. The dimension of acoustic model input feature vectors is 2,152 for the first scenario (H) and 2,973 for the second (I).

In both scenarios, several configurations are tested for each language. Because the acoustic model can be "controlled" by the speaker and gender identifying input features, it is interesting to see how these features affect the synthesis quality. We test the following combinations: speaker and gender features unset (default, D), set to the highest quality female speaker (EN-US, F)), highest quality male speaker (EN-GB, M), speaker of the "closest" language (C). In addition, for the second scenario, where we have the training data available for the language, we also test setting the speaker and gender features for this speaker (S).

Finally, for those languages for which we have speaker-specific data, we test the best-performing multilingual configurations from the above experiments (Best H and Best I) against acoustic models bootstrapped from speaker-specific data. The Vocaine vocoder [15] requires a gender-specific configuration, which currently defaults to females. We also use this default configuration for male speakers, which may negatively affect the vocoding quality.

For each subjective Mean Opinion Score (MOS) listening test we used 100 sentences not included in the training data for evaluation. Each rater was a native speaker of a language being tested and had to evaluate a maximum of 100 stimuli. Each item was required to have at least 5 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 for each language included at least 8 raters. Evaluation results are discussed next.

3.3. Results and Discussion

Table 1 shows the results of subjective listening tests for 12 languages not seen during the training of the ("held-out") acoustic model H (with 31 languages). For each language four configurations corresponding to different speaker/gender pairs described in the previous section were tested. The most related language in the training set shown in the table is denoted C, shown alongside the number of raters participating in the experiment. No closest speaker/gender evaluation was conducted for Dravidian languages (Tamil and Telugu) since these have no related languages in the training set. For Lithuanian (Baltic), Polish was chosen as the most related language even though these languages are not mutually intelligible [30]. As can be seen from the best results (shown in bold), there is no single winning speaker/gender feature combination across different configurations. Female gender is always preferred to male. On the scale between "poor" and "fair", the best scores for all languages lean towards "fair" (> 2.5). The intelligibility and naturalness of five out of 12 languages is rated as fair (> 3.0), with Bengali being best out of the pack.

Table 2 shows MOS results (along with confidence intervals) for six languages which were included in the training of the combined ("inclusive") acoustic model I (with 37 languages). Out of 12 languages under investigation (Table 1),

Table 1: *Subjective Mean Opinion Scores (MOS) (along with confidence intervals) for languages in various configurations for acoustic model (H) trained without these languages. The superscripts denote speaker and gender feature combinations: speaker/gender feature unset (D), best female speaker (F), best male speaker (M), speaker/gender of the "closest" language (C).*

Language	Family	Code (L)	H^D	H^F	H^M	H^C	C	Raters
Bengali	Indo-Aryan	BN	–	3.83±0.10	3.45±0.10	**3.91±0.10**	Hindi	14
Marathi	Indo-Aryan	MR	**3.16±0.12**	2.95±0.12	2.73±0.12	3.13±0.12	Hindi	9
Tamil	Dravidian	TA	2.78±0.08	**2.89±0.08**	2.68±0.07	–	–	10
Telugu	Dravidian	TE	**2.53±0.11**	2.50±0.11	2.35±0.10	–	–	11
Estonian	Finno-Ugric	ET	2.55±0.09	2.51±0.08	2.17±0.09	**2.55±0.09**	Finnish	11
Icelandic	Germanic	IS	**2.78±0.06**	2.69±0.05	2.30±0.08	2.72±0.07	Norwegian	9
Lithuanian	Baltic	LT	2.48±0.05	2.35±0.05	2.39±0.05	**2.50±0.06**	Polish	10
Romanian	Romance	RO	2.69±0.09	**2.80±0.09**	2.74±0.09	2.70±0.09	Italian	16
Serbian	Slavic	SR	3.21±0.05	**3.25±0.05**	2.92±0.05	3.23±0.04	Russian	10
Slovak	Slavic	SK	**2.71±0.08**	2.39±0.07	2.21±0.10	2.66±0.09	Czech	8
Slovenian	Slavic	SL	**3.15±0.10**	3.10±0.11	2.90±0.11	3.10±0.10	Russian	9
Ukrainian	Slavic	UK	3.08±0.09	2.72±0.08	2.83±0.09	**3.17±0.09**	Russian	16

Table 2: *Subjective Mean Opinion Scores (MOS) for languages in various configurations for acoustic model (I) that includes these languages. The superscripts denote speaker and gender feature combinations: speaker/gender feature unset (D), best female speaker (F), best male speaker (M), speaker/gender of the "closest" language (C), speaker/gender of this language (S).*

Code (L)	I^D	I^F	I^M	I^C	I^S	C
BN	–	–	**4.06±0.09**	3.88±0.10	–	Hindi
ET	2.69±0.09	2.53±0.09	2.55±0.08	2.71±0.10	**2.72±0.09**	Finnish
IS	**2.86±0.05**	2.79±0.06	2.80±0.05	–	–	–
LT	2.54±0.06	2.44±0.06	–	**2.67±0.06**	2.52±0.06	Polish
SK	**3.85±0.07**	3.68±0.07	2.60±0.09	3.70±0.06	3.84±0.06	Czech
UK	**3.48±0.10**	3.48±0.10	3.17±0.09	3.41±0.09	3.41±0.09	Russian

Table 3: *Best MOS scores for held-out multilingual AM (H), inclusive multilingual AM (I) vs. single-language AMs (`Single`) (with a source O).*

L	Best H	Best I	`Single`	O
BN	3.91±0.10	**4.06±0.09**	3.63±0.09	Goog
MR	**3.16±0.12**	–	1.26±0.07	IIIT
TA	**2.89±0.08**	–	2.65±0.08	IIIT
TE	2.53±0.11	–	**3.26±0.13**	IIIT
ET	2.55±0.09	2.72±0.09	**3.78±0.12**	Goog
IS	2.78±0.06	2.86±0.05	**3.07±0.07**	Goog
LT	2.50±0.06	**2.67±0.06**	1.85±0.08	Goog
SK	2.71±0.08	3.85±0.07	**4.05±0.07**	Goog
UK	3.17±0.09	3.48±0.10	**3.75±0.11**	Goog

at the time of writing there is no training data for Romanian, Serbian and Slovenian. The training data for Marathi, Tamil and Telugu from the lower audio quality IIIT corpus [31] was recorded at 16 kHz and was not included in the training of the combined 22.05 kHz model. Since Bengali and Icelandic are multi-speaker datasets, there is no unique speaker associated with them (hence I^S is left blank). Results for the rest of speaker/gender feature combinations are shown for each language. Best results are shown in bold. The scores of three out of six languages is in the "fair" range (> 3.2). Because Bengali is a multi-speaker male database, setting the gender feature to `male` seems to yield the best score (≈ 4.0). Pooling the data from other languages does not seem to dramatically improve Estonian, Icelandic and Lithuanian, even though the results lean from "poor" to "fair" (> 2.5).

Table 3 shows three-way comparison between the best performing configurations obtained with the "held-out" (H) acoustic model, combined acoustic model (I) and single-speaker LSTM-RNN system (`Single`) bootstrapped from the single-speaker database (denoted O). Best scores are shown in bold. For Bengali, Marathi, Tamil and Lithuanian, the multilingual acoustic model outperforms a single-speaker acoustic model. The improvement is most evident in case of Lithuanian, for which only a comparatively small database of around 1,000 utterances was available, not enough samples for training decent LSTM system. The huge improvement of Marathi over the IIIT system, as well as improvement of Tamil, can be attributed to the presence of Hindi to which both Marathi and Tamil are phonetically close. We attribute comparatively poor results from the "disappointing" batch of languages (e. g., Icelandic, Estonian, Slovak, Telugu and Ukrainian) to our suboptimial phonological features - the amount of linguistic sharing between these languages and other related languages in our database needs improving.

4. Conclusions

This paper investigated a multilingual acoustic modeling approach featuring a language- and speaker-agnostic model topology and a universal phonemic feature set. Experiments were conducted on twelve low-resource languages from diverse language families. We've shown that in two cases using an acoustic model with languages held out during training is better than building a dedicated single-speaker system from a small dataset. For two other cases, joint training with other languages also improves over a single-speaker system. These results support the hypothesis that the proposed approach may in certain situations benefit languages with small amounts of training data or no data at all. We also show that the complete absence of any training data may not necessarily lead to bad intelligibility – for all the languages in the held-out experiment the MOS scores are > 2.5 (anchoring the rating scale may be required in future work), while five of them display fair naturalness (> 3.0).

5. Acknowledgments

The author thanks Richard Sproat, Martin Jansche, Rob Clark, Alyson Pitts, Bo Li, Heiga Zen and anonymous reviewers for many useful suggestions.

6. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*. Vancouver, Canada: IEEE, May 2013, pp. 7962–7966.

[3] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks." in *Proc. of Interspeech*. Singapore: ISCA, September 2014, pp. 1964–1968.

[4] H. Zen and H. Sak, "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)*. Brisbane, Australia: IEEE, April 2015, pp. 4470–4474.

[5] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[6] P. Lanchantin, M. J. F. Gales, S. King, and J. Yamagishi, "Multiple-average-voice-based speech synthesis," in *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy: IEEE, May 2014, pp. 285–289.

[7] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Proc. of Interspeech*. Dresden, Germany: ISCA, September 2015, pp. 879–883.

[8] H. T. Luong, S. Takaki, S. Kim, and J. Yamagishi, "A DNN-based text-to-speech synthesis system using speaker, gender, and age codes," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2962–2962, 2016.

[9] A. Gutkin, L. Ha, M. Jansche, O. Kjartansson, K. Pipatsrisawat, and R. Sproat, "Building Statistical Parametric Multi-speaker Synthesis for Bangladeshi Bangla," in *SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages, 09-12 May 2016, Yogyakarta, Indonesia; Procedia Computer Science*. Elsevier, May 2016, pp. 194–200, edited by S. Sakti, M. Adriani, A. Purwarianti, L. Besacier, E. Castelli and P. Nocera.

[10] J. C. Paolillo and A. Das, "Evaluating language statistics: The ethnologue and beyond," *Contract report for UNESCO Institute for Statistics*, 2006.

[11] M. Versteegh, R. Thiolliere, T. Schatz, X.-N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The zero resource speech challenge 2015." in *Proc. of Interspeech*. Dresden, Germany: ISCA, September 2015, pp. 3169–3173.

[12] H. Zen, N. Braunschweiler, S. Buchholz, M. J. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1713–1724, 2012.

[13] Q. Yu, P. Liu, Z. Wu, S. Kang, H. Meng, and L. Cai, "Learning cross-lingual information with multilingual BLSTM for speech synthesis of low-resource languages," in *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE, March 2016, pp. 5545–5549.

[14] B. Li and H. Zen, "Multi-Language Multi-Speaker Acoustic Modeling for LSTM-RNN based Statistical Parametric Speech Synthesis," in *Proc. of Interspeech*. San Francisco: ISCA, September 2016, pp. 2468–2472.

[15] Y. Agiomyrgiannakis, "VOCAINE the vocoder and applications in speech synthesis," in *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*. Brisbane, Australia: IEEE, April 2015, pp. 4230–4234.

[16] International Phonetic Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[17] R. Jakobson, C. G. Fant, and M. Halle, "Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates." Acoustics Laboratoryn, MIT, Tech. Rep. 13, 1952.

[18] R. Jakobson and M. Halle, *Fundamentals of language*. Walter de Gruyter, 2002, vol. 1.

[19] A. Phillips and M. Davis, "BCP 47 - Tags for Identifying Languages," *IETF Trust*, 2009.

[20] M. P. Lewis, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the world*. SIL International, Dallas, TX, 2009, vol. 16.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks." in *Proc. of Interspeech*. Singapore: ISCA, September 2014, pp. 2268–2272.

[23] C. Ding, L. Xie, J. Yan, W. Zhang, and Y. Liu, "Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 98–102.

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

[25] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. Hinton, "On rectified linear units for speech processing," in *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*. Vancouver, Canada: IEEE, May 2013, pp. 3517–3521.

[26] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 137–140.

[27] K. Yu and S. Young, "Continuous F0 modeling for HMM based statistical parametric speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1071–1079, 2011.

[28] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *Proc. of Interspeech*. Singapore: ISCA, September 2014, pp. 338–342.

[29] Y. Miao, J. Li, Y. Wang, S.-X. Zhang, and Y. Gong, "Simplifying long short-term memory acoustic models for fast training and decoding," in *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE, March 2016, pp. 2284–2288.

[30] A. Lyovin, B. Kessler, and W. Leben, *Introduction to the languages of the world*. Oxford University Press, 2016.

[31] K. Prahallad, N. K. Elluru, V. Keri, S. Rajendran, and A. W. Black, "The IIIT-H Indic Speech Databases." in *Proc. of Interspeech*. Portland, Oregon: ISCA, September 2012, pp. 2546–2549.