# Areal and Phylogenetic Features for Multilingual Speech Synthesis

*Alexander Gutkin*[1], *Richard Sproat*[2]

[1]Google, London, United Kingdom
[2]Google, New York City, NY, USA

agutkin@google.com, rws@google.com

## Abstract

We introduce phylogenetic and areal language features to the domain of multilingual text-to-speech synthesis. Intuitively, enriching the existing universal phonetic features with cross-lingual shared representations should benefit the multilingual acoustic models and help to address issues like data scarcity for low-resource languages. We investigate these representations using the acoustic models based on long short-term memory recurrent neural networks. Subjective evaluations conducted on eight languages from diverse language families show that sometimes phylogenetic and areal representations lead to significant multilingual synthesis quality improvements. To help better leverage these novel features, improving the baseline phonetic representation may be necessary.

**Index Terms**: speech synthesis, neural networks, features

## 1. Introduction

With the advent of statistical parametric speech synthesis there has been an increase in research into multilingual acoustic modeling [1–5]. One of the main reasons for this is the rising demand for *polyglot speech synthesis* and speech-to-speech translation [6, 7]. In addition, it has recently been shown that pooling the data across multiple languages may improve overall synthesis quality [5].

One of the big research challenges facing the speech community is the growing need to support low and even zero-resource languages [8, 9]. For speech synthesis, this problem manifests itself in the lack of suitable high-quality training data and inadequate linguistic resources. One possible approach to tackle this resource scarcity problem is to build multilingual acoustic models, in the hopes that the presence of large amounts of high-quality training data from resource-rich languages will positively affect the synthesis of low-resource languages. In natural language processing, this type of multilingual joint learning has been shown to to be beneficial for tasks like morphosyntactic tagging [10].

An important aspect of building a multilingual model is the design of a shared linguistic representation for many diverse languages. More often than not, the multilingual linguistic feature representation is confined to phonetic transcription (in some universal format [11], to ease sharing between languages), the corresponding language-universal phonological features (such as place of articulation [12]) and basic language, region, speaker and gender-specific identifying codes. This choice of features often works well in practice, but may be insufficient for modeling similarity between diverse languages.

This paper investigates two types of typological information that may benefit cross-lingual modeling. Each type represents a different view of the languages. The *phylogenetic* features model the traditional approach to classification of languages placing the languages in the same cluster if they are demonstrably genetically related by descent according to historical reconstruction. The *areal* features, on the other hand, represent the languages by their geographic proximity to one another. Languages which, according to the phylogenetic view, reside in different classification trees, may end up in the same cluster if considered from an areal perspective. These types of features, though in a different form from our representation, are available in language typology and universals databases [13, 14]. However, to the best of our knowledge, this work is the first treatment of how areal and phylogenetic features affect multilingual speech synthesis.

We test the usefulness of the proposed features by constructing statistical parametric speech synthesis systems based on long short-term memory (LSTM) recurrent neural networks (RNNs). Aside from the multilingual input features and the larger number of parameters, the acoustic model is similar to the single-language single-speaker model proposed by Zen and Sak [15]. This paper is organized as follows: we introduce the proposed features in Section 2, followed by a brief outline of the architecture in Section 3. Experiments are described in detail in Section 4. We conclude the paper in Section 5.

## 2. Multilingual Linguistic Features

The core of our representation consists of phonetic features. The training set includes data from various languages and dialects, each following its own phonetic transcription convention. In order to train an acoustic model on such diverse multilingual data, we must transform each of the single-speaker phonemic configurations into a unified canonical representation using International Phonetic Alphabet (IPA) [11], similar to [3]. At present, this process is quite involved because it requires linguistic expertise to construct the mappings for each of the individual languages. Additional difficulties present themselves when these mappings disagree due to differences between transcribers, divergent transcription conventions, or the lack of native speakers to guide the design. For example, /ɐ̃ũ/ 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. In addition, each phoneme in our representation is decomposed into distinctive phonological features, such as place and manner of articulation [12].

### 2.1. Phylogenetic features

We use language and region identifying features based on the BCP-47 standard [16] to model the similarity between regional varieties 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.

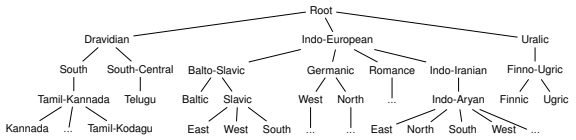


Figure 1: *Some language family trees represented up to four levels in the hierarchy. The root node is unused and is shown for convenience.*

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 [17] that represents related clusters of languages down to a depth of four levels. Figure 1 shows some of the phylogenetic features for the languages found in our corpus. 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. Areal features

The notion of a *linguistic area* was introduced by Emeneau [18], where he defines it (p. 16, fn. 28) as "an area which includes languages belonging to more than one language family but showing traits in common which are found to belong to the other members of (at least) one of the families", and notes that it is a translation of the earlier German term *Sprachbund*, due to Trubetzkoy [19]. Emeneau's particular focus was on India, home to three major language families and a number of smaller families. He focused on phonetic as well as morphological features that were shared across India regardless of language family, but for our purposes it is the phonetic features that are most relevant. He notes, for example (p. 7) that "most of the languages of India, of no matter which major language family, have a set of retroflex, cerebral, or dorsal consonants in contrast to dentals," and that "Indo-Aryan, Dravidian, Munda, and even the far northern Burushaski, form a practically solid bloc characterized by this phonological feature." These distinctions are most likely Dravidian in origin, but they spread to Indo-Aryan languages so that they are present in even the earliest forms of Vedic Sanskrit — striking since this is *not* a feature of Proto-Indo-European.

The relevance of areal features to the problem at hand is obvious. If, for example, one had no speech data for Marathi and wanted to build a synthesizer that sounds as much as possible like Marathi using data from other languages, it is probably more relevant to find data from languages of India, than from genetically related Indo-European languages of Europe, whose sound systems are rather different. Areal features are even potentially useful for Indian English, which inherits retroflex consonants from substrate languages [20] — making it distinct from dialects of English spoken elsewhere.

In our research, areal features depend upon a database of geographical information on the language. The basic information, which we currently have for 94 languages, includes the latitude ($\phi$) and longitude ($\lambda$) coordinate information that roughly defines the location ($l$) where the language in its current form evolved. Obviously this makes more or less sense depending on how widespread the language is today. So for Amharic, we place the location at Addis Ababa, and represent the information $(l, \phi, \lambda)$ as $(\texttt{Addis Ababa}, 9.0249700, 38.7468900)$, where the latitude and longitude are expressed in decimal degrees. While one can dispute the accuracy of this, it at least gives a roughly correct placement of the language on the globe. For English we place it around London, the source of Standard British English: $(\texttt{London}, 51.507351, -0.127758)$. This is obviously less reasonable for English, and indeed for our Indian English experiments we must replace this location with one more appropriate to an Indian language such as Hindi: $(\texttt{Delhi}, 28.613939, 77.209021)$.

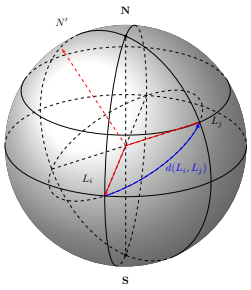


Figure 2: *Representation of two languages $L_i$ and $L_j$ as unit vectors on a sphere (shown in red). Distance $d(L_i, L_j)$ between them is defined along an arc (shown in blue).*

Figure 2 shows schematic representation of two languages $L_i$ and $L_j$ using simplified (spherical) Earth model. Given the latitude and longitude coordinates $(\phi_i, \lambda_i)$ for language $L_i$, there are several ways to represent this language in terms of acoustic model input features: First, one can represent the language by converting the lat-long coordinates to a corresponding three-dimensional *unit vector* (U, also known as $n$-vector) in spherical coordinates [21]. Second, each language $L_i$ can be represented in terms of its *distances* (D) to the rest of the languages $L_j$ in the set, $1 \leq j \leq N$, where the distance $d(L_i, L_j)$ is defined in terms of the length of an arc between the corresponding unit vectors. Finally, each language can be represented by *$n$-closest list* (N), which is a set of $n$ categorical features (given by BCP-47 language codes [16]) representing the $n$ languages closest (in terms of distance $d$) to the given language $L_i$.

## 3. System Architecture

The proposed architecture consists of multiple language-specific linguistic front-ends interfacing with a single acoustic back-end. The task of each front-end is text normalization, which involves converting the unnormalized text into detailed and unambiguous linguistic feature representation [22]. This consists of such basic tasks as tokenizing the text, splitting off punctuation, classifying the tokens and deciding how to verbalize non-standard words, i.e. things like numerical expressions, letter sequences, dates, times, measure and currency expressions [23].

The standalone back-end component of the architecture consists of a single acoustic model followed by the vocoder [24]. The input to the back-end is a set of linguistic features. The output is streamed audio (linear PCM). Unlike the systems recently reported in the literature [4, 5], we employ a simple network architecture similar to a single-speaker system. This is because it is hoped that the choice of the compact input feature space (described in Section 2) is general enough to pool data from related languages and yet accurate enough to discriminate between various languages and speakers. Moreover, supporting a new language in this system only involves

implementing a corresponding (possibly very basic) front-end but does not require re-training the acoustic model.

We use LSTM-RNNs designed to model temporal sequences and long-term dependencies between them [25]. These types of models have been shown to work well in speech synthesis applications [15, 26–28]. Our architecture is very similar to one proposed by Zen *et al.* [15, 29]: 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 vocoder 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 [15] is in the number of units in the rectified linear unit (ReLU) [30] 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 Section 4.

# 4. Experiments

## 4.1. System details

The multilingual corpus used for training the acoustic models has over 900 hours of audio and consists of 39 distinct languages. Some languages, such as English, comprise multiple datasets corresponding to different regional accents. For some regional accents (like `en-US`) we have several speakers. The corpus has both male and female speakers and is quite mixed: Most are single-speaker, while others, like Bangladeshi Bengali and Icelandic, have recordings from multiple speakers [31]. 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 speakers varies from low $F_0$ males to high $F_0$ females. No speaker normalization was performed on the data.

The speech data was downsampled to 22.05 kHz. Then mel-cepstral coefficients [32], logarithmic fundamental frequency ($\log F_0$) values (interpolated in the unvoiced regions), voiced/unvoiced decision (boolean value) [33], and 7-band aperiodicities were extracted every 5 ms, similar to [29]. These values form the output features for the acoustic LSTM-RNN and serve as input vocoder parameters [24]. 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 [29] 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 [24].

The architecture of the acoustic LSTM-RNN consists of $2 \times 512$ ReLU layers [30] followed by $3 \times 512$-cell LSTMP layers [34] with 256 recurrent projection units and a linear recurrent output layer [15]. The architecture of the duration LSTM-RNN consists of $1 \times 512$ 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 [35]. The duration LSTM-RNN was trained using an $\epsilon$-contaminated Gaussian loss function [29], whereas for acoustic LSTM-RNN the $L_2$ loss function was used because we observed it to lead to better convergence rates.

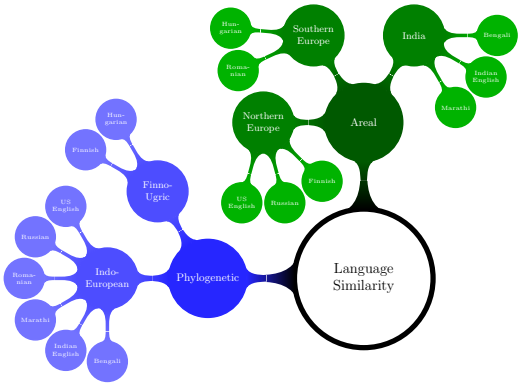


Figure 3: *Two possible views on representing language similarity: Areal (shown in green) and phylogenetic (shown in blue).*

## 4.2. Languages and model configurations

Eight languages were chosen for the experiments. Bengali, Finnish, Hungarian, Marathi, Romanian, Russian, Serbian and English. For English we selected both Indian and American dialects. We are interested in testing the hypothesis that the presence of areal and phylogenetic features influences the synthesis quality. Figure 3 shows how these languages are clustered under areal (shown in green) or phylogenetic (shown in blue) representation. Note that American and Indian English end up in different trees in the areal representation.

Marathi (Indo-Aryan language) and Serbian (Slavic) are "held-out" experiments, since we have no training data in 22.05 kHz for these language. It is interesting to see how well the system does when language has not been observed in training. For Bengali (Indo-Aryan language), we have small amounts of multi-speaker training data from [31]. We check whether the presence of Hindi (Indo-Aryan) and Indian English (West Germanic) in the training data affects these languages through areal and phylogenetic features.

Similar to the languages from an Indic area, we check the Romanian (Romance language geographically close to Hungarian and Serbian), Hungarian (Finno-Ugric language genetically related to Finnish and geographically close to Romanian) and Russian and Finnish (Slavic and Finno-Ugric languages geographically close to each other). Finally, American English should share information with Indian English. During synthesis, the speaker ID and gender input features are left unspecified for all languages apart from English.

We built eight acoustic model configurations each corresponding to a particular combination of input features. Each configuration, along with the types of the input features that comprise the input space, the dimensions of the input space and corresponding number of acoustic model parameters, is shown in Table 1. The baseline model (B) is trained on the input features that consist of IPA phonetic transcriptions along with distinctive phonological features, basic BCP-47 language/region tags and gender/speaker identifying features. The rest of configurations are obtained by augmenting the baseline input features with phylogenetic (Section 2.1) features (G) and/or areal

Table 1: *Input feature space makeup of various acoustic model configurations along with dimensions and number of parameters.*

| Configuration | Baseline | Genetic | Unit vector | Distances | $n$-closest | Duration AM | | Speech AM | |
|---|---|---|---|---|---|---|---|---|---|
| | B | G | U | D | N | # Inputs | # Params | # Inputs | # Params |
| B | ✓ | × | × | × | × | 1,598 | 2,394,626 | 1,602 | 4,325,589 |
| B+G | ✓ | ✓ | × | × | × | 1,665 | 2,428,930 | 1,669 | 4,359,893 |
| B+U | ✓ | × | ✓ | × | × | 1,601 | 2,396,162 | 1,605 | 4,327,125 |
| B+D | ✓ | × | × | ✓ | × | 1,688 | 2,440,706 | 1,692 | 4,371,669 |
| B+N | ✓ | × | × | × | ✓ | 1,749 | 2,471,938 | 1,753 | 4,402,901 |
| B+U+D | ✓ | × | ✓ | ✓ | × | 1,691 | 2,442,242 | 1,695 | 4,373,205 |
| B+G+U+D | ✓ | ✓ | ✓ | ✓ | × | 1,762 | 2,478,594 | 1,766 | 4,409,557 |
| B+G+U+D+N | ✓ | ✓ | ✓ | ✓ | ✓ | 1,913 | 2,555,906 | 1,917 | 4,486,869 |

Table 2: *Subjective Mean Opinion Scores (MOS) (along with 95% confidence intervals) for languages synthesized with various acoustic model configurations. Best scores are underlined. Statistically significant improvements shown in bold.*

| Language | B | B+G | B+U | B+D | B+N | B+U+D | B+G+U+D | B+G+U+D+N | Raters |
|---|---|---|---|---|---|---|---|---|---|
| Bengali | 3.66±0.07 | 3.70±0.07 | 3.60±0.07 | 3.72±0.07 | 3.67±0.07 | 3.70±0.06 | 3.66±0.07 | 3.70±0.07 | 22 |
| Finnish | 3.83±0.05 | 3.78±0.05 | 3.80±0.05 | 3.80±0.05 | 3.80±0.05 | 3.75±0.05 | 3.75±0.06 | 3.80±0.05 | 20 |
| Hungarian | 3.51±0.05 | 3.61±0.05 | 3.57±0.06 | 3.49±0.06 | 3.53±0.06 | 3.52±0.06 | 3.56±0.05 | 3.55±0.06 | 20 |
| English (IN) | 3.86±0.09 | 3.71±0.10 | 3.70±0.11 | 3.60±0.09 | 3.68±0.11 | 3.62±0.12 | 3.56±0.11 | 3.70±0.10 | $> 25$ |
| English (US) | 3.43±0.10 | 3.42±0.10 | 3.47±0.10 | 3.47±0.10 | 3.49±0.10 | 3.48±0.11 | 3.42±0.11 | 3.48±0.12 | $> 25$ |
| Marathi | 3.34±0.13 | 3.21±0.13 | 3.24±0.13 | 3.09±0.12 | 3.23±0.12 | 3.25±0.12 | 3.10±0.13 | 2.92±0.11 | 12 |
| Romanian | 3.13±0.06 | 3.26±0.05 | 3.08±0.06 | 3.06±0.06 | 3.26±0.06 | 3.05±0.06 | **3.39±0.06** | 3.35±0.06 | 23 |
| Russian | 3.12±0.10 | 3.05±0.09 | 3.15±0.10 | 3.07±0.10 | 3.08±0.10 | 3.11±0.09 | 3.06±0.09 | 3.11±0.10 | $> 25$ |
| Serbian | 2.75±0.06 | 2.82±0.06 | 2.83±0.06 | 2.74±0.06 | 2.80±0.07 | 2.74±0.06 | 2.77±0.06 | 2.77±0.06 | 9 |

(Section 2.2) features (U, D, N). For $n$-closest feature we used $n = 5$. As can be seen from the table, extending the input feature space with additional features does not dramatically increase the footprint of the resulting acoustic model. The difference between the simplest (B) and the most complex configuration (B+G+U+D+N) is approximately 16% increase in the number of features and the corresponding 0.04% increase in the number of model parameters.

Each configuration was evaluated using 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 8 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. For each language, all configurations were evaluated in a single experiment.

### 4.3. Subjective evaluation results and discussion

Table 2 shows the results of subjective listening tests for 9 languages, where for each language 8 acoustic model configurations described in the previous section were tested. Each mean opinion score is shown along with the corresponding 95% confidence interval [36]. The highest scores are underlined. The best configurations which exhibit no overlap in confidence intervals are deemed statistically significant and shown in bold.

Six out of nine languages exhibit slightly higher scores (shown underlined) for non-baseline configurations. For Indian English, Marathi and Finnish, the baseline configuration exhibits the highest score. However, due to the high overlap in 95% confidence intervals, for eight languages none of these differences are statistically significant. The only language which demonstrates significant improvements is Romanian. The best configurations (B+G+U+D and B+G+U+D+N) have the respective 0.26 and 0.22 improvement in mean opinin scores over the baseline. It is also interesting to note that there is no clear "winning" combination of phylogenetic and/or areal features across languages in our experiments. In addition, there does not seem to be on obvious correlation between the best configuration and the amount of data or the number of speakers for a particular language.

The lack of clear improvement for languages other than Romanian can possibly be explained by the following factor: Our phonetic transcription is rather sparse. Out of approximately 474 phonemes from 39 languages, there is a long tail of 250 phonemes that are only used in one language. Hence, there is not enough sharing of phonemes between the languages which will otherwise benefit from it (such as Bengali and Indian English). As a result, the phonetic features become strongly decorrelated. Consequently, phylogenetic and areal features are "too weak" to force a bond upon the overall representation of similar languages.

## 5. Conclusions

We introduced two novel types of linguistic features for training the multilingual parametric acoustic models for text-to-speech synthesis: areal and phylogenetic features. Although intuitively, such features should have a positive contribution to the overall synthesis quality, we showed that such claim is at present inconclusive. Out of diverse set of nine languages we were able to positively confirm this hypothesis for one language only (Romanian).

The above results, despite being promising, indicate that in our experiments the areal and phylogenetic features were "weaker" compared to the typical features used by the baseline system (phonetic transcriptions, language, region and speaker-identifying features). This warrants a thorough study into making the baseline phonetic feature space less sparse and trying alternative neural network representations for the acoustic model (such as [37]).

## 6. Acknowledgments

The authors thank Heiga Zen, Martin Jansche and the anonymous reviewers for many useful suggestions.

# 7. References


[1] A. W. Black and K. A. Lenzo, "Multilingual text-to-speech synthesis," in *Proc. of Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3. IEEE, 2004, pp. 757–761.

[2] C.-P. Chen, Y.-C. Huang, C.-H. Wu, and K.-D. Lee, "Polyglot speech synthesis based on cross-lingual frame selection using auditory and articulatory features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1558–1570, 2014.

[3] 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.

[4] 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.

[5] 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.

[6] K. Hashimoto, J. Yamagishi, W. Byrne, S. King, and K. Tokuda, "Impacts of machine translation and speech synthesis on speech-to-speech translation," *Speech Communication*, vol. 54, no. 7, pp. 857–866, 2012.

[7] S. Matsuda, X. Hu, Y. Shiga, H. Kashioka, C. Hori, K. Yasuda, H. Okuma, M. Uchiyama, E. Sumita, H. Kawai, and S. Nakamura, "Multilingual Speech-to-Speech Translation System: VoiceTra," in *Proc. of 14th International Conference on Mobile Data Management, Volume 2*. Milan, Italy: IEEE, June 2013, pp. 229–233.

[8] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[9] 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.

[10] H. O'Horan, Y. Berzak, I. Vulić, R. Reichart, and A. Korhonen, "Survey on the Use of Typological Information in Natural Language Processing," *arXiv preprint arXiv:1610.03349*, 2016.

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

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

[13] S. Moran, D. McCloy, and R. Wright, "PHOIBLE online," *Leipzig: Max Planck Institute for Evolutionary Anthropology*, 2014, `http://phoible.org`.

[14] P. Littel, D. R. Mortensen, and L. Levin, "URIEL Typological Database," *Pittsburgh: CMU*, 2016, `http://www.cs.cmu.edu/~dmortens/uriel.html`.

[15] 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.

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

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

[18] M. Emeneau, "India as a linguistic area," *Language*, vol. 32, no. 1, pp. 3–16, 1956.

[19] N. Trubetzkoy, "Proposition 16," in *Actes du Premier Congrès International des Linguistes*, 1928, pp. 17–18.

[20] Wikipedia, "Indian English," http://en.wikipedia.org/wiki/Indian_English#Consonants.

[21] C. Jekeli, "Geometric reference systems in geodesy," *Division of Geodesy and Geospatial Science, School of Earth Sciences, Ohio State University*, vol. 25, 2006.

[22] P. Ebden and R. Sproat, "The Kestrel TTS text normalization system," *Natural Language Engineering*, vol. 21, no. 03, pp. 333–353, 2015.

[23] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, "Normalization of non-standard words," *Computer Speech & Language*, vol. 15, no. 3, pp. 287–333, 2001.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] A. Gutkin, L. Ha, M. Jansche, K. Pipatsrisawat, and R. Sproat, "TTS for Low Resource Languages: A Bangla Synthesizer," in *Proc. 10th Language Resources and Evaluation Conference (LREC)*, Portorož, Slovenia, May 2016, pp. 2005–2010.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.