MEASURING THE EFFECT OF LINGUISTIC RESOURCES ON PROSODY MODELING FOR SPEECH SYNTHESIS

Andrew Rosenberg, Raul Fernandez, Bhuvana Ramabhadran

IBM Research AI
TJ Watson Research Center, Yorktown Heights, NY – USA
{amrosenb,fernanra,bhuvana}@us.ibm.com

ABSTRACT

The generation of natural and expressive prosodic contours is an important component of a text-to-speech (TTS) system which, in most classical architectures, relies on the existence of a text-analysis processor that can extract prosody-predictive features and pass them to a statistical learning model. These features can range from basic properties of the input string to rich high-level features which may not be always available when developing a TTS system in a new language with sparse computational resources. In this work we investigate how the prosody model of a speech-synthesis system performs as a function of different predictive feature sets that assume access to a certain amount of rich resources. We investigate, using objective metrics, the effect of relaxing the assumptions on input representations for prosody prediction for 5 languages, and evaluate the perceptual implications for US English.

Index Terms— prosody prediction, speech synthesis, low resources

1. INTRODUCTION

The generation of appropriate prosodic contours is one of the fundamental tasks that a Text-to-Speech (TTS) system needs to address to produce outputs that are perceived as natural, expressive, and consistent with the text. In classical unit-selection and parametric architectures, a natural-language text processor is responsible for analyzing an input string, and extracting a variety of predictors that can be given to a statistical model to learn (during training) and later generate (at run time) the necessary prosodic targets required by the architecture in question.

The extraction of these textual, prosody-predictive features is, to a large extent, a language-dependent task which incorporates knowledge about the factors that help shape the prosodic realization of utterances in the target language. Such knowledge can be incorporated into a system explicitly by a domain expert crafting linguistic rules, or automatically learned in a statistical framework from a variety of resources carrying linguistic knowledge. Examples of such resources include lexica annotated with phonetic baseforms and part-of-speech (POS) tags, syntactical or semantical treebanks, and corpora bearing symbolic prosodic annotations like prominence and phrasing.

The development of a TTS system in a new language for which no text-analysis processor is available will be initially constrained by the availability of these resources and/or access to linguistic knowledge. We would like to investigate, therefore, the question of how a prosody model performs as a function of variable predictive feature sets that assume access to a certain amount of rich resources (Section 2). This notion of resource-richness does not necessarily relate to the amount of textual data available in that language, even in an electronic format. A language may have good repositories of textual data, but poorly understood or studied intonational systems, and few or non-existing computational resources of the type already mentioned, both of which would hinder the development of rich linguistic features upon which to build prosody models. In this work, we investigate this topic by looking at the predictive ability of increasingly rich feature sets in various languages in the context of different machine-learning models for this task.

Interest in the utility of specific linguistic features to predict prosody has remained a topic relevant to speech synthesis in relatively high-resource languages for many years. Recently, more research attention has been given to speech synthesis in low-resource languages, where we lack, say, sufficient amounts of syntactically-annotated data to develop a computational resource, let alone a POS tagger or syntactic parser (Section 3).

Simultaneous to our investigation of the value of linguistic features with high resource requirements, we also compare the state of the art, bidirectional LSTM prosody assignment model to a new modeling strategy drawn from the neural machine-translation literature (Section 4). Our interest here is two-fold: can overall performance be improved, and is one of these approaches more or less robust to the impact of the differing feature sets?

.

2. LINGUISTIC REPRESENTATIONS

We consider the following nested sets of predictive features to study the impact on performance. The design of these sets has been motivated by sensible real-world assumptions about access to computational resources of variable richness.

Basic Set: In this most basic representation, we assume we have access to punctuation and knowledge of the language's phonetic inventory (and basic broad descriptions of the elements of that set like phonological voicing and vowel-consonant distinctions). In terms of prosodic structure, we assume access to sentence- and word-level tokenizations of the input string[1]. Regarding computational resources, we assume a lexicon is available to train a grapheme-to-phoneme (G2P) converter.

Medium Set: In addition to all the features derivable from the assumptions in the basic set, this medium representation assumes access to lexical stress annotation, and syllabification.

Rich Set: This set imposes the strongest assumptions on existing resources, and includes all those features derived in the medium set, in addition to POS tags, and symbolic word-level prominence. Symbolic prosodic structure is further assumed to include phrasing information. Note that, compared incrementally to the medium set, this is where we are relying on the heaviest resources and processors to obtain the prosody-predictive features.

After segregating the features by the types described above, we obtain the feature dimensionality summarized on Table 1 for each of five languages considered: US English, Castilian Spanish, Standard High German, French, and Korean. The numbers are roughly comparable across languages though differences exist due to the different cardinality of symbolic sets in various languages (e.g., number of POS tags in German vs. Spanish). This table also includes the size of each of the corpora used in the evaluation.

The specific features included in each set consist of categorical and numerically-defined features. The former include the different values taken on by a symbolic feature (e.g., various POS tags), to which one-of-N encoding has been applied, and the latter consist mostly of raw and normalized counts keeping track of elements between various boundaries (e.g., number of phones to the {previous,next} {syllable,word,phrase} boundary, etc.)

3. PREVIOUS AND RELATED WORK

There has been recent interest in the speech-synthesis literature on the challenges of building systems for under-resourced languages, addressing various aspects, to cite a few, like phrasing prediction [1], leveraging text-processing modules across closely-related but unequally-resourced languages, or data selection for improved intelligibility [2]. One main project worth singling out is the $Simple^4All$ Project [3], which had the expressed objective of facilitating the creation of speech technologies with little supervision, such as TTS front-end text-processing tools that made few implicit assumptions about the target language [4]

There is not an extensive literature investigating the role and effectiveness of different feature types on prosody prediction for TTS, but some relevant work has tried to investigate the effect of including richer representations on prosody modeling, demonstrating, e.g., the utility of employing syntactical information on top of segmental features for continuous targest prediction [5], or for symbolic prosodic phrasing assignment [6]. Our work departs from these lines of inquiry in that it explicitly seeks to segregate and quantify the effect of various types of information as a function of access to the resources needed to develop feature extractors that exploit this information. Although most of the cited work on synthesis for low-resource languages assumes a parametric framework, we anchor our exploration within a unit-selection synthesis framework, which relies on, and can fall back on, the inherent natural prosody of the selected units (see [7] for a discussion of the different roles prosody plays within these architectures). The case study of interest here is that of a new target language for which we assume a synthesis corpus of reasonable size can be obtained for building a voice, but for which we may lack the corresponding computational development resources for text processing and therefore prosody modeling.

Table 1. *Input feature dimensionality for each of the feature types*

Voice	Basic	Medium	Rich
English – Female (22.6 hrs)	156	202	302
Spanish – Male (9.5 hrs)	141	170	238
German – Female (13.8 hrs)	165	221	310
French – Female (22.1 hrs)	138	183	242
Korean – Female (17.1 hrs)	141	187	260

4. MODELING APPROACHES

In this work we highlight two neural-network-based modeling approaches: bidirectional LSTMs, a state-of-the-art baseline, and CBHG networks, a combined model that was developed for machine translation and has been used in end-to-end speech synthesis. In the preparation of this work, we performed preliminary investigations of a number of alternative architectures including basic CNNs, maxout-CNNs [8] and Self-Attention Networks [9]. These each yielded objective performance measures that were close to, but worse than, the

[1]For all the languages considered, and many others, word segmentation is straightforward, though this is not universally true.

CBHG and LSTM performance using sensible, but untuned settings. We did not perform exhaustive hyperparameter tuning on these other approaches, focusing our attention on the CBHG.

Bidirectional LSTMs: The Bidirectional LSTM approach we use was described in [7]. We have found that a 3 layer network with sizes, 65, 55, 45, is an optimal network structure. It is worth noting that this was tuned on US English and using the **Rich** feature set.

CBHG Network: The CBHG Network comprises a number of components. These, most critically, include a **C**onvolutional **B**ank at with filters of different time resolutions, a set of **H**ighway layers and a bidirectional **G**RU layer. The specific configuration we use is as follows. This configuration is consistent, regardless of the language or feature set that is used. The features are first processed by a prenet, which consists of a single fully-connected embedding layer with 128 hidden units, followed by a layer with 64 and another of 128 units. The pre-net is connected to a set of 6 1-D ReLU convolutional layers each with 128 filters, a stride of 1, and filter widths that vary from 1 to 6 units in time and a 1-D max-pooling layer of size 2. The output of the convolutional bank then is fed through two 1-D ReLU convolutional projection layers with 128 filters of size 3 with batch norm applied after each. A residual connection from the output of the prenet is applied to the output of the convolutional projections. This residual connection is followed by a single ReLU highway layer (again of size 128). Finally the output of the residual connection is fed into a single layer bidirectional GRU of size 128. A linear readout layer is used to convert its output to a 4 dimensional prosodic target vector. The prenet is trained with dropout with $p = 0.5$. A schematic of this network can be seen in Figure 1.

This network structure has been shown to be effective at capturing the long term dependencies necessary for machine translation [10] and speech synthesis [11].

5. EVALUATION

Models using the two modeling architectures previously described were trained for each of the five voices using a 90%-10% split of the data for training and validation, respectively. The prosody targets correspond to those used in a unit-selection system previously described in [7], and consist of a 4-valued vector specifying, for each unit, the unit's duration, initial and final f_0 values, and energy. Models were trained to minimize a weighted mean-square error (WMSE) criterion, where the target-specific weights were selected as follows. Duration: a weight of 1 for all speech units and utterance-medial silences, and 0 otherwise. Initial and final f_0: a weight corresponding to the unit's degree of voicing. Energy: a weight of 1 for all speech units, and 0 for all silences. During training, the log of each target was taken, and all targets were normalized to 0-mean and unit-variance

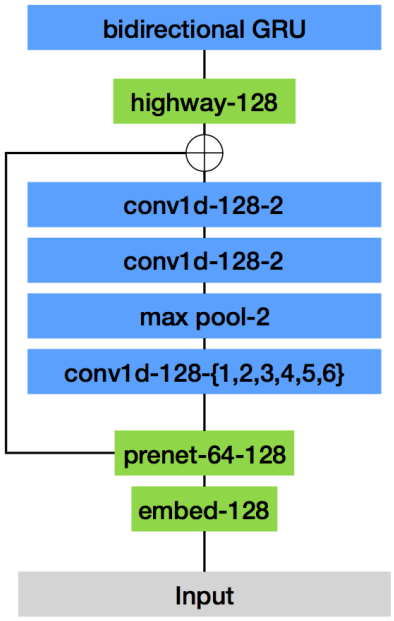

Fig. 1. *Diagram of the CBHG architecture.*

based on statistics of the training data.

Table 2 summarizes the WMSE on the validation set for the different voices as a function of the input feature sets. There is a clear trend across languages and models showing a degradation in the fit (i.e., higher loss) as we move from the richer set of features to the basic set. One exception to this trend is found in our Korean voice. Here we see similar prosody prediction across feature sets. There are a number of possible explanations for this. First, the overall error is much lower for Korean, the prosody in this voice may simply be easier to predict making differences between linguistic features to be insignificant. Second, the quality of the linguistic resources may be too low to impact prosody prediction. Third, the relationship between the specific linguistic

Table 2. *WMSE for various languages, architectures, and feature sets.*

Voice	Model	Basic	Medium	Rich
English	BiRNN	.433	.429	.419
	CBHG	.429	.424	.413
Spanish	BiRNN	.393	.389	.386
	CBHG	.408	.385	.386
German	BiRNN	.444	.438	.433
	CBHG	.437	.435	.430
French	BiRNN	.580	.577	.575
	CBHG	.575	.573	.571
Korean	BiRNN	.278	.279	.277
	CBHG	.274	.276	.278

resources that were explored and prosodic realization may be different in Korean than in the other languages.

To investigate the perceptual implication of this degradation, we conducted a series of listening tests using the US English voice by synthesizing, under various model-and-feature-set configurations, a set of 40 distinct sentences, 10 of which corresponded to syntactically-determined questions (i.e., non-declarative questions). Listeners in a crowd-sourced experiment were asked to rate the overall naturalness of a randomized subsample of the stimuli on a standard 5-point scale (Bad, Poor, Fair, Good, Excellent). Each sentence in each testing configuration was rated by 25 independent listeners.

Tables 3 shows the results of the mean opinion scores across the various feature sets under the BiRNN/LSTM and CBHG architectures. Although the degradation in the objective metrics corresponds to a gradual degradation in the perceptual quality of the systems, the only statistically-significant perceptual differences are between the basic and rich feature sets ($p = 0.0024$ for the BiRNN-LSTM and $p = 0.00017$ for the CBHG model), where these differences have been assessed using the Mann-Whitney U-Test adjusting for rater- and utterance-bias, as described in [12].

Table 3. *Overall MOS scores (and standard deviation) for the various combinations of architecture and feature sets.*

Architecture	Basic	Medium	Rich
Bi-RNN	3.36 (.95)	3.43 (.94)	3.47 (.93)
CBHG	3.33 (.95)	3.42 (.94)	3.48 (.96)

A second listening test was designed to isolate the perceptual effect of the alternative CBHG architecture against the BiRNN/LSTM baseline. A 2-system test using the same sentences, and the full rich feature set, was deployed using a similar design as previously explained and another 25 independent subjects. The results of that test show that the marginally-better fit of the CBHG architecture only translates into a statistically-insignificant marginal lead: an MOS of 3.45 ($\sigma = 0.94$) for BiLSTM vs 3.49 ($\sigma = 0.93$) for CBHG ($p = 0.295$). While we find a clear relationship between the objective criterion and MOS scores in all evaluations, substantial improvements to WMSE are required to observe statistically significant differences to MOS tests with 25 raters.

6. DISCUSSION AND CONCLUSION

Feature composition: We confirm that performance, as measured by objective metrics, degrades as a function of impoverished feature sets across the five languages examined. When evaluating how these differences translate into a perceptual difference for a US English voice, we notice that big improvements in objective scores are needed to obtain a measurable difference in MOS, and only observe a statistically significant difference between the basic and rich sets. There are two implications from this last observation worth noting: 1) a good amount of text analysis (e.g., parsing and tagging) is required to obtain a perceptually-measurable difference beyond the base quality level afforded by the coarser, but more easily accessible, features of the basic set. 2) On the other hand, when developing models for a new language where these resources are not immediately available, the results obtained with the medium set of features suggest it is possible to build acceptable prosodic-target models that provide a reasonable initial level of performance.

Modeling approaches: Since BiRNN-LSTM were first shown to provide state-of-the-art performance for prosody modeling tasks, many other neural architectures have been proposed in the deep-learning literature. This provided the impetus for reexamining their performance side by side with some of these newer models. Our experiments show that the BiRNN architecture still provides state-of-the-art performance, and that although CBHG is a viable competitor, it is a much more complex model that only yields small gains to objective measures and statistically-insignificant gains to subjective measures. Lastly, neither approach robustly compensates for the composition of the feature set, since we see the same degradation behavior across models as a function of input features.

We have reported initial perceptual validations for one language, but further work remains to be done to verify if similar perceptual trends accompany the reported objectives for the remaining languages. Another aspect that is not addressed in this work is the impact of the quality and quantity of respective resources on prosody prediction. For instance, we assumed access to a G2P and a POS tagger for various representations, but did not investigate variable factors like the impact of lexicon size on G2P performance, or the quality of the POS tags, and consequently the effect that such variable factors would carry over to the final task. That would have significantly increased the scope of this initial investigation. This is nonetheless, a reasonable inquiry (e.g., lexica for a new language may be of moderate size, or noisy; solving the G2P task will be a much more challenging problem for some languages; etc.), and it remains the object of further work.

7. REFERENCES

[1] A. Parlikar and A. Black, "Data-driven phrasing for speech synthesis in low-resource languages," in *Proc. ICASSP*, Kyoto, Japan, 2012, pp. 4013–4016.

[2] E. Cooper, X. Wang, A. Chang, Y. Levitan, and J. Hirschberg, "Utterance selection for optimizing intelligibility of TTS voices trained on ASR data," in *Proc. Interspeech*, Stockholm, Sweden, 2017, pp. 3971–3975.

[3] "Simple4all," `http://simple4all.org`, 2014.

[4] A. Suni, T. Raitio, T. Gowdha, R. Karhila, M. Gibson,

and O. Watts, "The Simple4All entry to the Blizzard Challenge 2014," in *Proc. Blizzard Challenge Workshop*, 2014.

[5] S. Hoffmann and B. Pfister, "Employing sentence structure: Syntax trees as prosody generators," in *Proc. Interspeech*, 2012, pp. 470–473.

[6] H. Che, Y. Li, J. Tao, and Z. Wen, "Investigating effect of rich syntactic features on Mandarin prosodic boundaries prediction," *J. Sign. Process Syst.*, vol. 82, pp. 263–271, 2016.

[7] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Using deep bidirectional recurrent neural networks for prosodic-target prediction in a unit-selection text-to-speech system," in *Proc. Interspeech*, Dresden, 2015, pp. 1606–1610.

[8] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. Laurent, Y. Bengio, and A.C. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," *CoRR*, vol. abs/1701.02720, 2017.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[10] J. Lee, K. aCho, and T. Hofmann, "Fully character-level neural machine translation without explicit segmentation," *CoRR*, vol. abs/1610.03017, 2016.

[11] Y. Wang, R.J. Skerry-Ryan, D. Stanton, Y. Wu, R.J. Weiss, N. Jaitly, Z. Yang, Y. Ying Xiao, Z. Chen, B. Bengio, Q.V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *CoRR*, vol. abs/1703.10135, 2017.

[12] A. Rosenberg and B. Ramabhadran, "Bias and statistical significance in evaluating speech synthesis with mean opinion scores," in *Interspeech*, August 2017, pp. 3976–3980.