# SUPERVISED AND UNSUPERVISED APPROACHES FOR CONTROLLING NARROW LEXICAL FOCUS IN SEQUENCE-TO-SEQUENCE SPEECH SYNTHESIS

Slava Shechtman<sup>1</sup>, Raul Fernandez<sup>2</sup>, David Haws<sup>2</sup>

# <sup>1</sup>IBM Haifa Research Lab, Haifa – Israel <sup>2</sup>IBM TJ Watson Research Lab, Yorktown Heights, NY – USA

slava@il.ibm.com, {fernanra, dhaws}@us.ibm.com

#### ABSTRACT

Although Sequence-to-Sequence (S2S) architectures have become state-of-the-art in speech synthesis, capable of generating outputs that approach the perceptual quality of natural samples, they are limited by a lack of flexibility when it comes to controlling the output. In this work we present a framework capable of controlling the prosodic output via a set of concise, interpretable, disentangled parameters. We apply this framework to the realization of emphatic lexical focus, proposing a variety of architectures designed to exploit different levels of supervision based on the availability of labeled resources. We evaluate these approaches via listening tests that demonstrate we are able to successfully realize controllable focus while maintaining the same, or higher, naturalness over an established baseline, and we explore how the different approaches compare when synthesizing in a target voice with or without labeled data.

*Index Terms*— prosody control, sequence-to-sequence speech synthesis

#### 1. INTRODUCTION

Sequence-to-Sequence (S2S) speech-synthesis architectures have become the state-of-the-art in the field, providing high-quality outputs that approach or match the perceived quality of natural speech in many studies. Aside from the level of quality attained, there are many attractive features to these models. They are able to jointly model different aspects of a waveform (e.g., segmental and prosodic), so interactions between them can be implicitly learned. They also do away with classical pipeline architectures in favor of a single unified model, which is appealing when some of the modules in the pipeline are difficult to develop (e.g., text-processing for a new language). On the other hand, they suffer from well-documented shortcomings, such as lack of interpretability (it can be difficult to tell which parts of the model are responsible for what functions), lack of controllability (it is more difficult to intervene into the model in order to control some aspects of the synthesis, which is often desired, such as when providing SSML support), and potential instability (small deviations at inference time can become exacerbated and generate highly degraded speech).

In this work, we address the *controllability* issue by expanding the S2S architecture with mechanisms that can be exposed to the user to manipulate some property of the output. Although usability factors are not the focus of this work, we nonetheless advocate for a set of properties that will make such controls accessible to the end consumer of the system, namely:

• Interpretability: The listener should be able to clearly hear and identify the effect of varying a control (e.g., speech is slower,

faster, higher-pitched, sounds happier, etc.).

- Monotonicity: A design that results in perceptual effects that vary monotonically as the user varies a control has a more intuitive feel, and is more easily tunable.
- Low-dimensionality: The user should not be expected to manipulate a large number of parameters to control the output. The model should either expose a low-dimensional controllable representation, or be able to step in and fill in defaults to obviate the task for the user.
- Disentanglement: Though this may be difficult due to the many ways different speech parameters interact, a set of controls that are more decoupled from each other facilitates the tuning of the output along fairly independent (perceptual) dimensions (e.g., tempo and volume could be tuned separately without needing to revisit a previously tuned parameter).

We explore the realization and controllability of narrow lexical focus as a case study for the above. Our objective is the realization of an emphatic level of prominence that is distinct from the type of accentuation that we observe in "neutral" broad-focus prosody. Consider the intonational phrase in the examples below when they occur as a reaction to the context in parentheses. In E1, as a reply to a general question, we see a likely case of broad focus prosody, where wine acts as the nuclear element and receives some sort of pitch accent. The same accented word, however, might be given a more emphatic degree of prominence when it happens in the context of E2. Furthermore, we can switch the focal point to a different word in the phrase when it is primed by a different context, as in E3. The [...] in these examples delimit the domain of focus, which the speaker may delineate, for instance, by employing a higher degree of disjuncture between the focal element and its context.

- E1: Mary is [pouring the wine]. (What's Mary doing?)
- E2: Mary is pouring the [wine]. (Is Mary pouring the beer?)
- E3: [Mary] is pouring the wine. (Is John pouring the wine?)

We are interested in the prosodic realizations that arise in examples such as E2 and E3 above (but also in other scenarios, such as contrastive emphasis, requesting clarification, etc.). In Sec. 2 we introduce an S2S architecture that supports this type of prosodic control, review in Sec. 3 how our approach compares to relevant research in the literature, evaluate competing approaches to this question in Sec. 4, and conclude in Sec. 5 with some analysis of these results and an outline of future steps.

#### 2. ARCHITECTURE

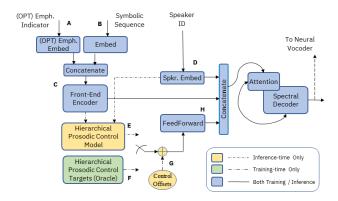
The model (Fig. 1) is a variant of the Tacotron2 architecture proposed in [1], augmented with components in the decoder to facili-

tate both the injection of controls, and improved stability during decoding [2]. This sequence-to-sequence model generates an acoustic spectral-prosodic representation that is then fed to an independently-trained, LPC-Net-based [3] neural vocoder to generate high-quality samples in real time [2].

The **Encoder** comprises the following components, combined as in Fig. 1 before being sent to the decoder:

- The emphasis embedding (A) from a Boolean indicator feature encoding emphatic focus within the utterance, as a way to provide direct supervision to the model.
- The *embedding of various linguistic symbols* (B) extracted from an extended phonetic dictionary comprising phone identity, lexical stress, phrase type, and other symbols for word boundaries and silences. This analysis is carried out externally by a rules-based TTS Front End module, adopted from a unit selection system [4].
- A front-end encoder (C) consisting of convolutional and bidirectional Long Short-Term Memory (Bi-LSTM) layers (as in [1]), encoding the merged embeddings from (A) and (B).
- A global utterance-level speaker embedding (D), broadcast over the length of the sequence, to support training in a multispeaker setting.
- A set of 4-dimensional hierarchical prosodic controls (which will be introduced in Sec. 2.1) designed to enable the type of fine, word-level modification needed to realize the prosodic patterns associated with emphatic focus. Since these prosodic controls are a set of statistics extracted from the acoustic signal, the ground-truth values from the training set are used during training (F). At inference time a separate predictive module (E) steps in to provide default predictions for the hierarchical prosodic trajectories.
- An optional *user-exposed control* (G) to modify the default predictions generated by (E). In particular, we propose a set of additive controls that are linguistically intuitive and interpretable (Sec. 2.1). Note that the feed-forward operation in block H is placed *after* the (optional) user request in G. This design choice is made to preserve the interpretability of the quantities the user gets to manipulate (which would not be the case if the order was reversed, and the independent prosodic targets were blended via a non-linear feed-forward operation).

The **Decoder** is an autoregressive network that largely follows the standard Tacotron2 architecture, but with modifications on the attention mechanism, autoregressive feedback, choice of targets, and training losses. These have already been described in [2] and are summarized here as follows. The attention is an augmented two-stage attention where the content- and location-based attention of Tacotron2 are followed by a structure-preserving mechanism encouraging monotonicity and unimodality in the alignment matrix. This modification has been found to be crucial to increase stability during inference, particularly in the presence of external controls. A double feedback approach is used during training to expose the model both to the previous ground-truth output value (i.e., teacher forcing) as well as the previous predicted value (i.e., inference mode). At inference time, the predicted value is replicated. The model is trained in a multi-task fashion to predict the 80-dim mel cepstral features in tandem with the parameters needed as inputs for an independently trained LPC-Net neural vocoder. For 22kHz signals, these features (which we denote as "LPC features") consist of a 22-dim vector with 20 cepstral coefficients,  $\log f_0$  and



**Fig. 1.** System architecture. The dashed line indicates the output of the S2S model is sent to a separately trained neural vocoder which does not play a role in the optimization of Eqn. 1.

 $f_0$  correlation. The predicted LPC features are also processed with two post-nets (one to refine the cepstrum, and one to refine the pitch parameters); no post-net refinement is applied on the mel task. Let  $y_t^M$  and  $y_t^L$  represent the target sequences for the mel and LPC tasks respectively,  $\tilde{y}_t^M$  and  $\tilde{y}_t^L$  their final predictions, and  $\hat{y}_t^L$  the "intermediate" LPC-feature prediction (before the post-net). Then the following differential loss function is used to train the system:

$$\mathcal{L} = MSE(\tilde{y}_t^M, y_t^M) + 0.8MSE(\hat{y}_t^L, y_t^L) + 0.4MSE(\tilde{y}_t^L, y_t^L) + 0.4MSE(\tilde{y}_t^L, y_t^L) + 0.4MSE(\Delta \tilde{y}_t^L, \Delta y_t^L), \tag{1}$$

where the  $\Delta$  operator applies the first difference in time to a sequence, and MSE(,) is the mean-squared error. For the sake of space, we omit some detail in this exposition, and refer the reader to [5, 2] for additional background and formulae.

This architecture accommodates some variants depending on the availability of labeled resources and types of control that are exposable to a user. Among them, we explore the following:

- Classic Supervision: When labeled data is available, this architecture conditions directly on a Boolean indicator feature. During training, the ground truth values of the audio signals are used. During inference a binary request is passed to the system<sup>1</sup>. This corresponds to blocks {A, B, C, D} on the encoder side.
- *No Supervision:* Under the assumption that *no* labeled data exists, the architecture defined by {B, C, D, E, F, G, H} provides a way to introduce sensitivity into the S2S system during training, and, at inference time, control the realization of the prosodic patterns via a *tunable* set of controls (cf. the binary control of the supervised architecture).
- Hybrid: Though components E through H are motivated by an
  unsupervised approach, they may facilitate the realization of
  prosodic patterns even when labeled data exists by working in
  tandem with an explicit feature. To investigate this, we consider a
  "hybrid" approach (defined by the full model {A-G}) that mixes
  supervised knowledge with the infrastructure designed to tackle
  the case when we don't have access to it.

<sup>&</sup>lt;sup>1</sup>This value could be either user-specified for given words (e.g., via markup) or inferred from text. We do not address here the problem of inference from text (though we have previously in [6]), and focus on the realization of prosodic controls assuming an existing request.

#### 2.1. Hierarchical Prosodic-Control Model

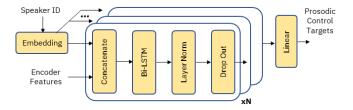
Following the motivation for a perceptually-interpretable, low-dimensional control mechanism for prosody discussed in Sec. 1, we propose a hierarchical set of four prosodic controls that summarize information about the duration and pitch excursion of a signal over linguistically-meaningful and intuitive intervals of the prosodic hierarchy. These controls include global and local properties, and are an extension of the approach in [5], which allowed for controlling global aspects like overall tempo, but which lacked any control to effect the kind of deviation from long-term trends needed to realize local emphatic focus. To arrive at these, let us first define the following statistics:

- S<sub>dur</sub>: The log of the average per-phone durations, along a sentence (and excluding any silence).
- $S_{f_0}$ : The log- $f_0$  "spread" (defined as the difference between the 95- and 5-percentiles of log- $f_0$ ), along a sentence.
- $W_{dur}$ : The log of the average per-phone durations (as above), along each word.
- $W_{f_0}$ : The log- $f_0$  "spread" (as above), along each word.

Note that the average per-phone durations in the above definitions are estimated as the duration of speech (in seconds) along the relevant spans (word or sentence) divided by the number of phone symbols contained therein, and that therefore no fine-level phonetic alignment is required in the computation (only coarse word-level alignments and either phonetic transcriptions or a dictionary). These sentence- and word-level properties are propagated down to the temporal granularity of the phonetic encoder outputs (i.e., phones) to form piecewise functions that are constant within a (sentence or word) unit. From this we define the following four-component prosodic-control target vector:

$$PC = Norm_{\sigma}\{[S_{dur}, S_{f_0}, W_{dur} - S_{dur}, W_{f_0} - S_{f_0}]\}, \quad (2)$$

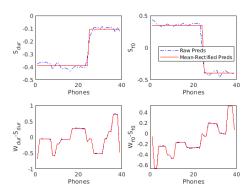
where  $Norm_{\sigma}\{\}$  is the linear map  $[-3\sigma^2, 3\sigma^2] \rightarrow [-1, 1]$ , and  $\sigma^2$  is the global (corpus-wide) variance for each of the statistics in PC. At inference time, the predictions of the prosodic-control subnet are rectified to be piecewise constant as the oracle values that the S2S system was trained with. In the evaluated systems, a mean pooling function is applied to the prediction to be constant between the (known) sentence and word boundaries.



**Fig. 2.** Architecture of the hierarchical prosodic sub-network for predicting targets from encoder-level features.

The architecture of the prosodic-control predictor (Fig. 2) consists of a stack of N blocks, each comprising a concatenation of the speaker embedding with the block's input, a Bi-LSTM, Layer normalization [7], and Drop-Out. Models are trained in a multi-speaker fashion via a speaker-embedding layer whose output is fed into every cascaded block. (We will discuss how we instantiate model sizes for the different components of this architecture when we discuss the

details of selecting models for evaluation in Sec. 4.) Since the replication to the phone level artificially introduces an over-contribution to the loss, each observation in each of the prosodic targets is downweighted by this replication factor (e.g., for the sentence-level targets, each phone-level observation in a 10-phone sentence receives a weight of 0.1; a similar approach is applied to the word-level targets). These observation-level weights (uniquely determined by prosodic constituency) are then combined with global target-specific weights  $\alpha$  that can be set during training to trade-off between the different targets (in this evaluation  $\alpha=[1,1,1.5,3.5]$ ). The model is then trained with ADAM [8] to minimize the weighted L1 loss between predictions and targets. A set of 10% of the sentences in the training set are held out to tune structure (e.g., number of hidden units and blocks) and learning rate hyper-parameters .

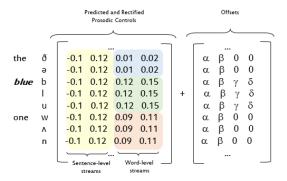


**Fig. 3.** Sample phone-level trajectories of the four prosodic controls for a two-sentence input.

At run time, lexical focus is controlled by the process illustrated in Fig. 4: The prosodic-control predictions generated by component E in Fig. 1, and post-processed to be piecewise constant, are offset by 4 tunable parameters  $(\alpha,\beta,\gamma,\delta)$  where the  $(\alpha,\beta)$  are global sentence-level offsets (that are applied uniformly and therefore only contribute to the overall expressiveness of the utterance) and  $(\gamma,\delta)$  boost the word-level predictions of  $\mathit{only}$  those words we wish to make salient (remaining non-focal words receive no offset). These run-time hyperparameters can be tuned via an independent development set.

### 3. PREVIOUS AND RELATED WORK

Synthesizing emphasis has been previously explored within other architectures like unit selection [9, 10, 11], classical parametric synthesis [12, 13, 14], and pipeline systems using neural networks [15]. Within S2S models, controllability has recently received a moderate amount of attention, with the Global Style Tokens (GST) proposal of [16] being one of the earliest works to discover latent styles in an unsupervised fashion. GST-based approaches have found wide usage (see, e.g., [17, 18, 19]), but as these representations are discovered, rather than explicitly formulated, they often lack a priori interpretability (though post hoc listening often reveals some uniform perceptual quality). GST and others [20] where global tempo is controllable lack the finer-grained level of control we pursue due to its global nature. Non-GST approaches include works like [21], where direct conditioning on estimated indicators of emotion are used to control the output. Recent work by [22, 23] has also looked at the controllability of prosodic properties in Transformer-based neural TTS systems, although at the core of that approach is a move away



**Fig. 4.** Boosting the prosodic-control predictions of sentence- and word-level targets to realize focus. The example shows a fragment of an utterance where the word *blue* is to be emphasized (e.g., *I don't want the red one; I want the blue one*). The predicted prosodic controls are offset by global and local offsets, where the local offsets are applied to the focal words only.

from S2S models that, for the sake of speed, replaces a S2S teacher with feed-forward student models that decouple prosodic from spectral modeling. We, in contrast, retain the full S2S framework in our implementation. Hierarchical representations and controllability have been explored together in [24, 25] though these approaches lack the level of interpretability for fine word-control. The work of [26] targets interpretable and controllable hierarchical prosodic controls and comes closest to the approach we pursue. However, their disentanglement is data-driven and leaves some residual couplings between the (pitch and duration) dimensions we control separately; we model  $f_0$  dynamics (as opposed to levels), which is more perceptually relevant to realizing emphatic focus; and as we will see in the next section, our controllable systems attain the same or higher level of quality when introducing this prosodic variation. Prosody transfer across databases bearing different labels is one of the main applications of our framework, as we will discuss in Sec. 4. The works of [27, 28] pursue similar goals although, being based on globalsentence level embeddings, they do not address fine-level control as we do.

## 4. EVALUATION

The training material comprised four corpora from three professional native speakers of US English, broken down as follows: a set of 10.8K sentences from a male speaker (M1); a set of 1K sentences from the same male speaker, where each sentence contains several emphasis-bearing words ( $M1_{emp}$ ); and two corpora from two distinct female speakers (F1 and F2) containing approximately 17.3K and 11K sentences respectively. The corpus  $M1_{emp}$  was collected by indicating to the speaker the emphasis-bearing words within each sentence, and instructing him to realize an emphatic level of prominence on those target words. His prosodic realizations differ in marked ways from the style of broad focus prosody in terms of tempo, relative pitch accent height, and disjuncture from adjacent material. The sentences were intended to serve as elicitors for various cases of narrow focus (e.g., contrast, disambiguation, etc.). Notice that labeled data is available for only one speaker, and that the size of this corpus is considerably smaller than that of the base corpora. A sentence from  $M1_{emp}$  contains three emphatic words on average, and the overall percentage of such words was approximately 23%. We define the following data partitions to facilitate the ensuing discussion: a set of data with all the resources pooled, including the emphatic data  $D_{emp} = \{M1, 10 \times M1_{emp}, F1, F2\}$ , and a base set  $D_{base} = \{M1, F1, F2\}$ . Note that  $D_{emp}$  uses 10-fold replication of the  $M1_{emp}$  subset to compensate for the lower prior.

We would like to investigate the trade-offs between approaches that use labeled data (when available), and the fully unsupervised approach that is possible within the framework proposed in Sec. 2. To that end, consider the following systems:

- Base (NoEmph): A baseline S2S system, which uses global (sentence-level) prosodic controls, but no word-level prosodic control. The training set (D<sub>emp</sub>) subsumes the emphatic data, but no other emphasis-marking feature is used.
- Base (Sup): A baseline system with Classic Supervision (as in Sec. 2) with global controls, trained with D<sub>emp</sub> and an explicit binary feature encoding the location of emphasis.
- PC-Unsup: A Fully Unsupervised system (as per Sec. 2) with variable prosodic control, where both the S2S and prosodyprediction components are trained with D<sub>base</sub>.
- PC-Hybrid: A Hybrid system with variable prosodic control, trained with D<sub>emp</sub>, and an explicit Boolean emphasis indicator as in the Baseline (Sup) model.

**Table 1**. Summary of the different properties and training strategies among the different systems evaluated.

	Base	Base	PC-	PC-
	(NoEmph)	(Sup)	Unsup	Hybrid
Control?	N	Y	Y	Y
Type of	None	Binary	Tunable	Binary /
Control?	None	Billary	Tullable	Tunable
Training?	$D_{emp}$	$D_{emp}$	$D_{base}$	$D_{emp}$
Emph. feat?	N	Y	N	Y

The architecture of **Base** (**NoEmph**) with global controls was already presented and evaluated in [5]. Since it lacks fine-grained lexical prosodic control, we do not expect it to perform well on an emphasis-evaluation task. It is used here, however, to provide a strong anchor point with respect to overall quality to ensure that the alternative proposals do not degrade with respect to the naturalness afforded by this approach. A common LPC-Net neural vocoder, also trained in a multi-speaker fashion using  $D_{base}$ , was used for all experiments [2].

Model selection and tuning was done as follows. First, for the prosodic sub-network, 10% of the training data was held out to do a grid search over structures and learning rate by tracking the heldout loss. The models thus selected were, for the PC-Unsup condition, a stack of 5 blocks with 175 hidden units in the Bi-LSTM layer, and, for the **Hybrid** model, a stack of 4 blocks with 200 hidden units in the Bi-LSTM layer. The speaker embedding was of dimension 20 in both cases. Once this was fixed, a development set of 20 sentences not used in training was used to perceptually tune remaining hyper-parameters of the different configurations, including the dimension of the emphasis-embedding space (dim = 8 for the Hybrid model, and 16 for the Base (Sup) system), and the runtime additive word-level boosting parameters  $(\gamma, \delta)$  (see the "control offset" component G in Fig. 1, and Fig. 4) for the PC-Unsup and **PC-Hybrid** word-level controls (set to (0.25, 1.30) and (0.0, 1.5)respectively). These word-level offsets were applied only to the item

in a sentence that was intended to be the focus carrier; the predictions of the prosodic-control model remain unboosted for all other lexical items. Sentence-level boosting was not found to provide any advantages over word-level boosting, and the parameters  $(\alpha,\beta)$  (see Fig. 4) were therefore only used for the two reference systems (Base (NoEmph) and Base (Sup)), and set to (0.0,0.5)). The non-negative boosting values we employ match our theoretical expectations, and what we empirically observe in the  $M1_{emp}$  subset, that focused items receive more pronounced pitch accents, and slower speaking rate/longer durations. We observed that the  $Base\ (Sup)$  system already realized these tempo differences quite well, and only boosted the pitch excursions when tuning the Hybrid systems. In general, we find that after tuning a single set of boosting parameters works quite well across a variety of sentences and voices  $^2$ .

### 4.1. Subjective Listening Tests

We wish to evaluate how the different multi-speaker approaches we have described fare in a perceptual listening task. In particular, we are interested in examining two test-case scenarios. In the first case, we operate under the assumption that the target synthesis voice matches a speaker for whom we have existing training data (i.e., the *matched* condition). In the second, and more interesting case, we assume that the target synthesis voice lacks any such labeled resources for training (though some exists for a separate speaker), and that therefore any use the system makes of supervised information is done indirectly by transferring knowledge from one speaker to another (we refer to this as the *transplant* condition). Notice that the distinction we have just introduced applies to the systems that are sensitive to supervision in some way (i.e., **Base** (Sup) and PC-Hybrid); system PC-Unsup, by construction, is not.

To evaluate the systems defined in the previous section, while addressing the *matched* and *transplant* cases respectively, we conducted two independent listening tests where the target speakers were M1 (whose training data contains an emphatic subset) and F1 (whose training data does not)<sup>3</sup>. No natural recordings were included (which could have provided a topline performance) since no common set of utterances with emphasis existed for both voices, and we wanted to run parallel tests. Instead, we opted for an evaluation set of 43 unseen sentences, with each containing a single focused word.

**Table 2.** MOS  $(\sigma)$  results for the matched condition (M1). For *emphasis* all systems are statistically significantly different from each other. For *quality*, there are no statistically significant differences between the pairs {Base (NoEmph), PC-Unsup} and {Base (Sup), PC-Hybrid}; all other pairwise differences are significant. Significance is assessed at the p=0.01 level via one-tailed t-tests.

System	Attribute		
System	Emph	Quality	
Base (NoEmph)	2.21 (1.3)	3.87 (0.8)	
Base (Sup)	4.08 (1.0)	4.10 (0.1)	
PC-Unsup	3.35 (1.2)	3.82 (0.9)	
PC-Hybrid	3.96 (1.0)	4.08 (0.8)	

The listening tests were designed to evaluate the systems in terms of two attributes on 5-point scales: (i) how well they realize

narrow focus on a given word, and (ii) the overall quality of the sentence. Listeners were recruited through a crowd-sourcing platform and presented with one audio sample at a time, accompanied by a transcript of the text where the intended focus-carrying word had been capitalized. To facilitate comprehension of the task we provided the listeners with the following set of instructions, and collected their responses in the provided 5-point scales:

The UPPERCASE word (excluding the word "I", if it exists) in the text above should sound emphasized in this sample. Assess the level of emphasis you hear in the UPPERCASE word. It sounds: 1 (neutrally spoken), 2, 3 (somewhat emphasized), 4, 5 (definitely emphasized). Assuming the UPPERCASE word is emphasized as requested, rate the overall quality and naturalness of this audio sample: 1 (Bad), 2 (Poor), 3 (Fair), 4 (Good), 5 (Excellent).

Each {sentence, system} combination received 25 independent rating tuples (one for each of the 2 attributes). The texts were designed to make the choice of focus semantically congruent with the context-providing sentence. Tables 2-3 summarize the restuls in terms of Mean Opinion Scores (MOS), standard deviation  $(\sigma)$ , and pairwise statistical significance.

**Table 3.** MOS  $(\sigma)$  results for the transplant condition (F1). All pairwise differences are statistically significantly different for *emphasis*. For *quality*, {Base (Sup.), PC-Unsup} are statistically equivalent; all other pairwise differences are statistically significantly different. Significance is assessed at the p=0.01 level via one-tailed t-tests.

System	Attribute		
System	Emph	Quality	
Base (NoEmph)	2.20 (1.3)	3.87 (0.9)	
Base (Sup)	3.71 (1.2)	3.97 (0.9)	
PC-Unsup	3.58 (1.1)	3.97 (0.9)	
PC-Hybrid	4.02 (1.0)	4.08 (0.8)	

### 5. DISCUSSION AND CONCLUSIONS

From these evaluations, we can make the following remarks for both speakers. All controllable systems achieved a much higher degree of emphasis than Base (NoEmph) (which, as expected, attained low scores in terms of emphasis realizability), and this was achieved at no expense of overall quality since the remaining systems are statistically better or the same. We hypothesize this improvement in quality is due to the fact that conditioning on additional prosodic attributes of the outputs steers the model toward more natural (and stable) points during training. We observe differences between the approaches, however, comparing the matched vs. transplant conditions: when labeled data is available for a target speaker, our experiments suggest that the fully-supervised approach offers the best operating point in terms of both quality and emphasis (Table 2). This approach, however, does not generalize as well as the hybrid approach does to a new target speaker lacking labeled data (Table 3). For the latter, combining supervision with the prosodic-conditioning framework supplements the performance for both attributes when training a multi-speaker model to enable the transfer of knowledge. Lastly, we see that even lacking any labeled data, the framework is able to provide a good point of quality and emphasis control by means of boosting the predictions of the fully unsupervised model. This is facilitated by our use of a set of controls that are readily interpretable and can be perceptually linked to the task at hand. Though the results are very encouraging, some difficult test cases remain.

 $<sup>^2</sup> Samples$  and additional listening test details are available at http://ibm.biz/SLT2021.

 $<sup>^{3}</sup>$ In informal listening, we found F1 and F2 to be of comparable quality, so only one voice was selected to keep the test manageable.

For instance, we have observed in informal listening the challenge posed by some function words, particularly clitics or words containing only unstressed vowels in broad-focus realizations.

We have introduced and validated a framework that allows for a finer degree of control over lexical prosody to guide the realization of narrow focus in S2S synthesis. This framework encompasses a set of user-driven controls that meet the criteria that we highlighted and advocated for in Sec. 1 of the paper: they consist of a low-dimensional representation of prosody, they are intuitive in the sense that changes to the controls map to identifiable perceptual effects in the output, and they offer a mechanism that disentangles different components of prosody (duration and pitch) that can be tuned separately. The approach requires only a moderate amount of knowledge external to the framework in the form of coarse word-level alignments, and we have shown that it can accommodate various degrees of supervision depending on available resources, with different variants bringing in different strengths depending on the operating conditions (e.g., synthesizing from a speaker with labeled supervised data vs. transplanting to a novel speaker that lacks such resources).

We should note that this framework can also be extended to include other levels of the prosodic hierarchy to explore expressive effects beyond localized narrow focus. For instance, incorporating the intonational phrase into the analysis might provide a way to better model the pitch reset associated with parentheticals. Addressing the shortcomings already mentioned and incorporating these extensions remain the subject of ongoing and future work.

#### 6. REFERENCES

- [1] J. Shen, R. R. Pang, R.J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R.A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 4779–4783.
- [2] S. Shechtman, R. Rabinovitz, A. Sorin, Z. Kons, and R. Hoory, "Controllable sequence-to-sequence neural TTS with LPC-NET backend for real-time speech synthesis on CPU," *CoRR*, 2020.
- [3] J. M. Valin and J. Skoglund, "LPCNET: Improving neural speech synthesis through linear prediction," in *ICASSP*, Brighton, England, 2019, pp. 5891–5895.
- [4] J. Pitrelli, R. Bakis, E.M. Eide, R. Fernandez, W. Hamza, and M.A. Picheny, "The IBM Expressive Text-to-Speech Synthesis System for American English.," *IEEE Trans. Audio, Speech and Lang. Processing*, vol. 14, no. 4, pp. 1099–1108, July 2006
- [5] S. Shechtman and A. Sorin, "Sequence to Sequence Neural Speech Synthesis with Prosody Modification Capabilities," in *Proc. SSW10*, Vienna, Austria, 2019, pp. 275–280.
- [6] Y. Mass, S. Shechtman, M. Mordechay, R. Hoory, O.S. Shalom, G. Lev, and D. Konopnicki, "Word emphasis prediction for expressive text to speech.," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 2868–2872.
- [7] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *CoRR*, vol. abs/1607.06450, 2016.
- [8] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, San Diego, May 2015.
- [9] A. Raux and A.W. Black, "A unit selection approach to f0 modeling and its application to emphasis," in *Proc. ASRU*, Saint Thomas, VI, 2003, pp. 700–705.

- [10] R. Fernandez and B. Ramabhadran, "Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis.," in *Proc. SSW6*, Bonn, Germany, 2007, pp. 34–39.
- [11] V. Strom, R. Nenkova, A.and Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, and D. Jurafsky, "Modeling prominence and emphasis improves unit-selection synthesis," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 1282–1285.
- [12] K. Yu, F. Mairesse, and S. young, "Word-level emphasis modelling in HMM-based speech synthesis," in *Proc. ICASSP*, Dallas, TX, 2010, pp. 4238–4241.
- [13] F. Meng, Z. Wu, H.M. Meng, J. Jia, and L. Cai, "Hierarchical english emphatic speech synthesis based on HMM with limited training data," in *Proc. Interspeech*, Portland, OR, 2012, pp. 466–469.
- [14] Q.T. Do, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A hybrid system for continuous word-level emphasis modeling based on HMM state clustering and adaptive training," in *Proc. Interspeech*, San Francisco, CA, 2016, pp. 3196–3200.
- [15] S. Shechtman and M. Mordechay, "Emphatic speech prosody prediction with deep LSTM networks," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 5119–5123.
- [16] Y. Wang, D. Stanton, Y. Zhang, R.J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R.A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *CoRR*, vol. abs/1803.09017, 2018.
- [17] R.J. Skerry-Ryan, E. Battenberg, Xiao. Y., Y. Wang, D. Stanton, J. Shor, R.J. Weiss, R. Clark, and R.A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," *CoRR*, vol. abs/1803.09047, 2018.
- [18] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *Proc. ICASSP*, Brighton, U.K., 2019, pp. 5911–5915.
- [19] J. Valle, R.and Li and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *Proc. ICASSP*, Barcelona, Spain, 2020, pp. 6189–6193.
- [20] J. Park, K. Han, Y. Jeong, and S.W. Lee, "Phonemic-level duration control using attention alignment for natural speech synthesis," in *Proc. ICASSP*, Brighton, U.K., 2019, pp. 5896– 5900.
- [21] X. Zhu, S. Yang, G. Yang, and L. Xie, "Controlling emotion strength with relative attribute for end-to-end speech synthesis," in *Proc. ASRU*, Singapore, 2019, pp. 192–199.
- [22] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T-Y. Liu, "FastSpeech: Fast, robust and controllable Text to Speech," in Advances in Neural Information Processing Systems 32, pp. 3171–3180. 2019.
- [23] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end Text to Speech," CoRR, 2020.
- [24] X. An, Y. Wang, S. Yang, Z. Ma, and L. Xie, "Learning hierarchical representations for expressive speaking style in endto-end speech synthesis," in *Proc. ASRU*, Singapore, 2019, pp. 184–191.

- [25] W-N. Hsu, Y. Zhang, R.J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," in *Proc. ICLR*, New Orleans, 2019.
- [26] G. Sun, Y. Zhang, R.J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchichal fine-grained prosody modeling for interpretable speech synthesis," in *Proc. ICASSP*, Barcelona, Spain, 2020, pp. 6264–6268.
- [27] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-grained robust prosody transfer for single-speaker neural text-to-speech," in *Proc. Interspeech*, Graz, Austria, 2019, pp. 4440–4444.
- [28] Y-J. Zhang, S. Pan, L. He, and Z-H. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *Proc. ICASSP*, Brighton, U.K., 2019, pp. 6945–6949.