VOICE-TRANSFORMATION-BASED DATA AUGMENTATION FOR PROSODIC CLASSIFICATION

Raul Fernandez[1], *Andrew Rosenberg*[1], *Alexander Sorin*[2], *Bhuvana Ramabhadran*[1], *Ron Hoory*[2]

[1]IBM TJ Watson Research Center, Yorktown Heights, NY – USA
[2]IBM Haifa Research Lab, Haifa – Israel

{fernanra,amrosenb}@us.ibm.com, sorin@il.ibm.com

ABSTRACT

In this work we explore data-augmentation techniques for the task of improving the performance of a supervised recurrent-neural-network classifier tasked with predicting prosodic-boundary and pitch-accent labels. The technique is based on applying voice transformations to the training data that modify the pitch baseline and range, as well as the vocal-tract and vocal-source characteristics of the speakers to generate further training examples. We demonstrate the validity of the approach by improving performance when the amount of base labeled examples is small (showing reductions in the range of 7%-12% for reduced-data conditions) as well as in terms of its generalization to speakers unseen in the training set (showing a relative reduction in the error rate of 8.74% and 4.75%, on the average, for boundaries and accent tasks respectively, in leave-one-speaker-out validation).

Index Terms— data augmentation, voice transformation, prosody labeling, recurrent neural networks

1. INTRODUCTION

The topic of automatic labeling of prosodic events has received considerable attention in the literature on prosodic analysis as corpora that carry annotations prominence and phrasing annotations can serve as a valuable resource for speech scientists. Examples of the types of application that such a resource could provide include enabling various linguistic analyses (e.g., studying intonation variation) or developing speech-language technologies (e.g., the prosodically-marked text associated with a speech corpus could be used when training a data-driven phrasing module in a text-to-speech front-end). Creating such databases by human annotation is known to be a notoriously laborious and expensive effort, a fact that has motivated automating the task in some way. Still, although a variety of modeling techniques have been proposed, most of the successful approaches can be situated within fully supervised or semi-supervised frameworks that rely on some amount of labeled data to train a classifier. Given this dependence on labeled data, we are then interested in approaches that will improve the performance for small or moderate amounts of training data since this represents a realistic scenario in terms of data acquisition. A related limitation brought over by small databases concerns the speaker variability observed in the corpora, and thus the ability of a model trained on this limited data to generalize outside the training pool. We are therefore also interested in approaches that will degrade gracefully when presented with unseen speakers.

Motivated by these observations, in this paper we explore techniques which augment the amount of labeled training examples without having to procure new independent (hand) annotations for the new training material. At the core of these data-augmentation (DA) approaches is an automatic transformation of the input exemplars in such a way that we can still associate them with the original labels; that is to say, the transformation is label-preserving with respect to a particular label of interest. In this work, we are interested in modeling binary prosodic labels indicating whether a word receives a pitch accent or is followed by an intonational phrase boundary, based on acoustic input features extracted from the audio stream. We propose to investigate a set of acoustic transformations that alter the (global) fundamental frequency, speaker's vocal tract, and voice quality (by voice source manipulation) in a way that preserves the local patterns of prominence and phrasing, and thus allows for the generation of many more training exemplars for training the model.

2. DATA AUGMENTATION VIA TRANSFORMATIONS OF THE VOCAL SOURCE AND TRACT

The approach we explore in this work relies on generating multiple "copies" of the acoustic training material that is distinct enough from the original data in feature space, yet respects the same sequence of categorical prosodic labels we want to model: pitch accents and intonational phrase boundaries. For this type of label-preserving audio modification, we use an analysis-resynthesis and voice-transformation framework to separate vocal-tract and glottal-excitation components. Although a detailed description of the approach lies beyond the scope of this paper, in what follows we review the fundamental aspects of the modeling to offer enough insight into the augmentation technique.

Analysis: At this stage, a pitch contour is first extracted from the audio signal at a 5ms-update rate. The audio signal is then analyzed at the same frame update rate. All unvoiced frames are skipped whereas each voiced frame is analyzed by a window containing 3.5 pitch cycles to yield glottal-source and vocal-tract parameters. The glottal source is represented by the Liljencrants-Fant (LF) glottal-source parametrization [1], represented by the 3-parameter vector $\theta = [T_p, T_e, T_a]^T$ (normalized by the pitch period), the aspiration noise level, and the gain factor. The vocal tract is represented by 40 Line Spectral Frequencies (LSF). The temporal trajectories of all parameters are smoothed with a 7-frames long moving averaging.

Reconstruction: When the audio signal is reconstructed from the parametric representation, consecutive voiced frames are stacked together to form contiguous voiced regions. These voiced regions are then interleaved with the unvoiced regions, which are kept in the raw PCM representation. A voiced region is synthesized as follows: First a sequence of consecutive pitch cycle onsets is generated according to a desired synthesis pitch contour, which may be either provided externally or derived from the original one. The glottal-

Table 1. *Summary of transformations.* f_0^{shift} *is specified in octaves (a shift of* ± 1 *corresponds to raising/lowering the baseline pitch by one octave) whereas* f_0^{range} *is multiplicative. The pairs in the vocal-tract map corresponds to inflection points (given here in kHz) of the interpolating spline. T1 through T3 share a common VT transform and apply different pitch and glottal transformations. T4 leaves the vocal tract unchanged, whereas T6 and T7 leave the glottal sources unchanged.*

ID	$(f_0^{shift}, f_0^{range})$	VT_{map}	β_{lt}	$\{\alpha, \theta_{ref}\}$
T_1	(-0.5,1)	$\{(1,1.1),(2,2),(3,2.8),(4,3.7)\}$	.7	-
T_2	(-0.1,1)	$\{(1,1.1),(2,2),(3,2.8),(4,3.7)\}$	.27	-
T_3	(-2.5,1.5)	$\{(1,1.1),(2,2),(3,2.8),(4,3.7)\}$	.24	-
T_4	(-0.4,2.4)	-	-	$\{.9,[.45,.8,.01]\}$
T_5	(-2,1)	$\{(.7,.62),(1.2,1),(2.3,2),(2.8,3),(3.7,3.9),(7,6.8),(9.3,9.5)\}$	.9	-
T_6	(0,1.1)	$\{(1,.85),(2,1.85),(3,2.75),(4,3.6)\}$	0	-
T_7	(0,1)	$\{(1.2,1),(2,2.05),(3,3.2),(3.85,4)\}$	0	-

source and vocal-tract parameters associated with each pitch cycle are generated by interpolating between the corresponding parameters associated with the cycle's surrounding (edge) frames. The sequence of glottal pulses is generated, and each pulse is multiplied by its corresponding gain factor. Additive aspiration noise is constructed for the entire voiced region by amplitude modulation of a 500-Hz high-passed Gaussian noise signal. The amplitude modulation forms the noise time-envelope shape, so that it is aligned with the glottal-pulse energy envelope, respects the noise level and gain values within each cycle, and evolves smoothly at the transitions between the consecutive cycles. The LSF parameters associated with each pitch cycle are converted into auto-regression filter coefficients. Finally, the glottal source undergoes a time-varying auto-regressive filtering where the filter coefficients are updated at the beginning of each pitch cycle, and each voiced region is then combined with its neighboring unvoiced regions using an overlap-add process.

Transformation: The previously described reconstruction algorithm provides the basis for introducing global (time-invariant) voice modifications that alter vocal tract, glottal pulse, pitch, and speech rate. Global pitch modifications are introduced by transposition and stretching of the original pitch contour by factors f_0^{shift} and f_0^{range} respectively. The vocal tract transformation takes the form of an interpolating spline function, with user-specified inflection points, that is used to map each cycle's LSFs prior to reconstruction. For the glottal pulse transformations we enable two independent types of control for modifying the cycle's glottal-parameter vector θ: either by (i) interpolating with user-provided reference glottal-pulse vector θ_{ref} and mixing weight $0 \leq \alpha \leq 1$:

$$\hat{\theta} = (1-\alpha)\theta + \alpha\theta_{ref} \tag{1}$$

or by (ii) interpolating between two stylized (pre-computed) pulses corresponding to lax and tense voice qualities:

$$\hat{\theta} = \begin{cases} (1-\beta_{lt})\theta + \beta_{lt}\theta_l & \text{if } \beta_{lt} > 0 \\ (1-|\beta_{lt}|)\theta + |\beta_{lt}|\theta_t & \text{otherwise} \end{cases} \tag{2}$$

where $-1 \leq \beta_{lt} \leq 1$ is a user-specified parameter that trades between lax and tense qualities (and recovers the original pulse when $\beta_{lt} = 0$), and $\theta_l = [.5, .9, .099]^T$ and $\theta_t = [.1, .15, .00001]^T$ are the stored lax and tense glottal parameters respectively.

We implemented 7 different transformations with which to augment the training data, summarized on Table 1. These were empirically chosen so as to provide a good amount of variability in terms of identity and expressiveness. Because of the global nature of the transformations, local contextual changes relevant to the perception of prominence and phrasing are preserved, a fact that was also informally tested by the authors before conducting any experiments.

3. PREVIOUS AND RELATED WORK

There is work in the prosodic-modeling literature that has previously addressed the issue of limited resources, with techniques such as semi-supervised learning [2, 3] and active learning [4, 5] being deployed to compensate for the lack of labeled examples when training prosodic classifiers. Separately, DA techniques have received recent focused attention, particularly in the deep-learning literature, given the renewed interest in building very large neural-network models that deliver substantial gains when trained with large corpora. The most relevant area to ours in which these techniques has been applied include the acoustic models (AM) of speech recognition systems, though some other applications, like language modeling (e.g., [6], where machine translation is used to enlarge a set of limited textual resources), or acoustic-event detection(e.g., [7], where acoustic signals sharing a common event description are mixed) have also benefited from similar methodology. Many of the speech techniques have in common the judicious introduction of variants, at some point in the processing pipeline, that are meaningful and consistent with some aspect of the speaker's speech-production process. Some of the instantiations of these types of acoustic DA ideas include altering the speed of speech [8] or the input feature space to the AM to correspond to vocal tract perturbations [9, 10], or to linear feature-space cross-speaker mappings [10]. Other speech-based DA approaches focus instead not on augmenting the speaker space, but on creating more channel noise variability to improve the robustness of the AM to unseen noise conditions [11].

The exploration of vocal-tract transformations is one of the aspects which our proposal shares with previously used techniques. The proposed approach, however, provides for a much richer set of transformations taking into account various manipulation related to vocal-source production that allow the creation of different voice identities preserving the same phonetic and prosodic sequences. This type of technique has yet to be explored within the DA literature. That exploration as a pure DA technique, and its particular application to the task of automatically classifying prosodic events, are, to the best of our knowledge, novel contributions of this work.

4. PROSODIC CLASSIFICATION

4.1. Features

A word-level vector of acoustic features is constructed using the AuToBI tool [12] to extract a series of measures found to have been effective in detecting either pitch accents or boundaries in previous work [13]. We assume knowledge of the start and end times of words, but do not use any other lexical information. In this work,

the same feature vector is common to both detection tasks, and contains features identical to those described in [14]. In the interest of space, we present a high-level description of the 1,915 features used, and refer the reader to that paper for additional detail.

Most acoustic features are constructed from short-time frame analysis contours aggregated over each word. The base contours are pitch (log Hz), intensity (db), the product of pitch and intensity, spectral tilt, fundamental frequency variation [15] and deltas of these contours. The aggregations include minimum, maximum, standard deviation, mean, and z-score of the maximum in the context of the word. Additional shape representations include the area under the curve, tilt coefficients, tonal center of gravity and likelihood of a curve being a rise, fall, peak or valley. We incorporate context by normalizing maximum and mean values within a word by statistics from windows over surrounding words. Features are also extracted from specific subword regions: the lexically stressed syllable and the final 200ms. The feature vector also contains features crafted to exploit the difference between changes in pitch and intensity. Additionally we extract presence and duration of preceding and following silence, the duration of the word, and rate of voicing.

The ground truth used to train all systems is derived from perceptual ToBI labels provided by expert annotators (AuToBI is only used as a front-end feature extractor). Prominence is modeled as a binary task resulting from collapsing all pitch-accent labels; phrase boundaries correspond to major (intonational) boundaries.

4.2. Network Structure and Training Recipe

We perform all experiments using bi-directional recurrent neural networks (BiRNN), as we have found these to provide state-of-the-art classification for these tasks in previous work [14]. All systems use an initial non-recurrent (dimensionality-reduction) projection layer, which feeds into bi-directional recurrent layers, all of which use gated recurrent units [16] as their non-linear activation functions. In early experiments, we found these to work as well or better than Long Short-Term Memory units [17], while using fewer parameters.

While both tasks use a similar structure, the two experiments differ in the hyperparameters which result in the best performance. To identify them, we use a grid search and evaluate performance on a held-out development set. The grid search explores the following parameters: 1) Size of the projection layer (15 or 25 for phrase detection, 25 or 50 for pitch accent detection), 2) Size of the recurrent layers (10, 20 or 30 units), 3) number of recurrent layers (2, 3, 4) , 4) initial learning rate (1e-3, 1e-4, 5e-5) , 5) momentum (.5, .75, .9) , and 6) magnitude of random weight initialization (.01, .05).

All RNNs are trained using *rmsprop* [18] to minimize a cross-entropy loss function, with gradient clipping set to 50, and initial learning rates and momentum set via the grid search. We use a batch size of 1, and evaluate performance against the development set after each iteration over the training data. We train until neither the loss function nor the classification error fail to decrease for 5 iterations, and evaluate the model with the best classification error.

Due to the sensitivity of NNs on initializations and order of data presentation [19], for each grid-search setting, we train three classifiers and use the median performance as the representative result for model selection. In the case of ties, we select the hyper-parameter setting with the smallest standard deviation across the three runs.

5. EXPERIMENT DESCRIPTIONS

We are interested in exploring whether DA can be of help in the case of limited data resources for supervised learning, and whether it can improve generalization to speakers for whom no labeled data is available during training. Both of these reflect realistic operating points, as we argued in the introduction. To explore these questions, we have constructed a few case scenarios which are now described in some detail. In what follows, the sets $\mathcal{T}, \mathcal{D}, \mathcal{S}$ denote, respectively, the training, development, and test sets that are used within the RNN training recipe described above, and for final testing. The different augmentation schemes below are applied to the training set; the dev and test sets are never augmented.

Case 1: Speaker-independent case. Here we explore how DA can improve generalization to a speaker not seen in the training set. For these experiments, we rely on the Boston University Radio News Corpus (BURNC) [20], which contains prosodic annotation for multiple (6) speakers. We split the (uneven) amount of data from each speaker into disjoint 90% and 10% training and development sets respectively, Denoting by $T_j\{Z\}$ the set of utterances built by applying any of the $1 \leq j \leq 7$ voice transformations described in Section 2 to an arbitrary set Z, we define the augmented training set $Z^{tr,aug}_{spkr} = \{Z^{tr}_{spkr}, \bigcup_{j=1}^{7} T_j\{Z^{tr}_{spkr}\}\}$ for each speaker. We then follow a leave-one-speaker-out testing-evaluation methodology with the following sets: $\mathcal{T}_s \doteq \{\bigcup_{n=1;n\neq s}^{6} Z^{tr,aug}_{spkr=n}\}$, $\mathcal{D}_s \doteq \{\bigcup_{n=1;n\neq s}^{6} Z^{dev}_{spkr=n}\}$, $\mathcal{S}_s \doteq Z_{spkr=s}$, and $s = \{1, \cdots, 6\}$ is a speaker ID.

Case 2: Speaker-dependent case. In this scenario we assume we have access to a fixed amount of labeled data from a single target speaker of interest and construct several experiments that rely on a corpus of recordings from a unit-selection speech-synthesis system, a portion of which (3700 utterances) has been annotated with pitch-accent and phrase-boundary labels. This corpus is initially split into disjoint training, development, and test subsets containing 2984, 373, and 373 utterances respectively (corresponding to a 80%–10%–10% partition). To investigate the data-size effect, the training and development sets are downsized to form smaller subsets with $k = \{5, 10, 20, 40, 80\}$ percent of the initial number of utterances contained in each. All the downsized sets form nested subsets, so $X^{tr}_{5\%} \subset X^{tr}_{10\%} \cdots \subset X^{tr}$ (similarly for the downsized development subsets). This is intended to mimic the case where more training labeled data becomes progressively available for the speaker. We then consider the following augmentation schemes.

Voice-transformation-based DA: For any training condition indexed by the amount of data k, we define the augmented training set as before $X^{tr,aug}_{k} = \{X^{tr}_{k}, \bigcup_{j=1}^{7} T_j\{X^{tr}_{k}\}\}$. Then $\mathcal{T}_k \doteq X^{tr,aug}_{k}$ and $\mathcal{D}_k \doteq X^{dev}_{k}$ are used as training and dev sets. To facilitate a direct comparison, a unique testing set $\mathcal{S} \doteq X^{test}$ (at its original size, with no DA) is kept as the target testing set for all k-sized conditions.

Use of additional labeled resources and no DA (Data Pooling): This case represents a simple pooling of resources to improve the performance of a target speaker, with no additional transformation of the original data. Although this is not a true DA technique in the sense meant in this paper, we are interested in contrasting the utility of augmenting in-domain data (the previous scheme) against simply adding independent, but possibly out-of-domain, resources with good-quality labels, when such resources are available (which may not always be the case). To address this scenario, we reuse the BURNC from Case 1 as the auxiliary (out-of-domain) corpus[1]

[1] Although both the synthesis corpus and BURNC contain carefully read speech, they show noticeable stylistic differences. They have also been collected under different recording conditions, sampled at different rates (22.05 vs 16kHz), and their prosodic annotations have been provided by different labelers.

to investigate the effect of data pooling as a function of training data size on the target speaker. (The auxiliary corpus is assumed to be of fixed size.) Recycling the previous notation, we define the following data-size-dependent training and developments sets: $\mathcal{T}_k \doteq \{X_k^{tr}, \bigcup_{s=1}^{6} Z_{spkr=s}\}$ and $\mathcal{D}_k \doteq X_k^{dev}$. As before, the test set remains $\mathcal{S} \doteq X^{test}$

Data pooling combined with data augmentation: In this final scheme we also assume access to an auxiliary labeled resource of fixed size, but exploit it within a DA framework to examine how it impacts the performance on a target speaker of interest as we vary the amount of in-domain training data (i.e., the previous 2 schemes combined). As before, we define the relevant learning sets as follows: $\mathcal{T}_k \doteq \{X_k^{tr,aug}, \bigcup_{s=1}^{6} Z_{spkr=s}^{tr,aug}\}$, $\mathcal{D}_k \doteq X_k^{dev}$, and $\mathcal{S} \doteq X^{test}$

6. RESULTS AND CONCLUSION

Tables 2 and 3 summarize the results of the various experiments previously outlined for the speaker-independent (SI) and -dependent (SD) cases. A few remarks are in order:

Table 2. *Summary of speaker-dependent experiments showing classification error (%) for baseline and augmented systems. Shown in parenthesis on the augmented column is also the relative error reduction (%), defined so that a positive sign (in bold) corresponds to a reduction with respect to the baseline. The top speaker column also shows the number of word tokens for each speaker. The average errors (WAvg) are weighted with respect to these counts.*

Task	Speaker	Baseline	Augmented
Phrase Bdry.	F1 (3,681)	7.39	**7.06 (4.47)**
	F2 (12,697)	7.54	**7.08 (6.10)**
	F3 (2,733)	6.99	**5.74 (17.88)**
	M1 (4,955)	7.09	**5.71 (19.46)**
	M2 (3.537)	7.63	**7.15 (6.29)**
	M3 (1,935)	4.55	4.70 (-3.30)
	WAvg.	7.21	**6.58 (8.74)**
Pitch Accent	F1	12.47	12.66 (-1.52)
	F2	14.54	**13.71 (5.71)**
	F3	13.50	**12.44 (7.85)**
	M1	13.65	**13.57 (0.59)**
	M2	12.61	**11.42 (9.44)**
	M3	12.71	**11.89 (6.45)**
	WAvg.	13.69	**13.04 (4.75)**

– We observe improvements for the two tasks considered under one or more of the data-augmentation schemes explored over a strong state-of-the-art baseline, which we have previously validated against other strong classification approaches [14].

– On the average, the proposed DA offers a substantial relative improvement in its generalization to speakers unseen in the training data. For each of the tasks, however, there is one speaker for whom the performance worsens, and we continue to explore ways in which the DA schemes and/or the training recipe can be made more robust to improve the worst-case performance.

– DP by itself, in addition to not always being feasible in terms of resource availability, is not as reliable as the proposed DA scheme (possibly due to domain missmatches). This is particularly noticeable for the boundary classification task, where DA never performs worse than the baseline, and can offer considerable reductions when the amount of training data is small (7.5-9.7% relative). We observe a much less consistent pattern in terms of the utility of the DP scheme. For the accent task, we see that DA can lead to a decreased performance with respect to the baseline, but in this case too, it provides a more consistent behavior than the DP scheme alone.

– Merging external resources and applying the transformation (DA+P) provides the best scheme, with better performance than either DA or DP alone. This is most notable in the full-data condition, where it substantially beats a state-of-the-art baseline performance to give us the best numbers we have obtained to date for this task and dataset. There is only one case in which this scheme fails to beat the baseline (pitch-accent classification with 60% of the data), but this may be due to the fact that the baseline attains, uncharacteristically, its best-performance at this point. (Notice that this point is an outlier with respect to the monotonic behavior we otherwise observe as a function of data size for the baseline accent task.)

Table 3. *Summary of speaker-dependent experiments showing classification error (%) for baseline and augmented systems using data-augmentation (DA), data-pooling (DP), and data-augmentation-and-pooling (DA+P) schemes. Shown in parenthesis is the relative error reduction (RER) (%), defined so that a positive sign (in bold) corresponds to a reduction with respect to the baseline.*

	Phrase Boundary Error (RER) (%)			
Size (%)	Base	DA	DP	DA+P
100	6.53	**6.51 (0.31)**	6.83 (-4.59)	**6.08 (6.89)**
80	6.85	**6.76 (1.31)**	6.92 (-1.02)	**6.56 (4.23)**
60	7.17	**6.69 (6.69)**	7.27 (-1.39)	**7.03 (1.95)**
40	7.76	**7.24 (6.70)**	**7.37 (5.03)**	**7.47 (3.74)**
20	8.05	**7.64 (5.09)**	8.69 (-7.95)	**7.97 (0.99)**
10	9.69	**8.75 (9.70)**	**9.03 (6.81)**	**8.29 (14.45)**
5	10.46	**9.68 (7.46)**	**9.17 (12.33)**	**9.80 (6.31)**
	Pitch Accent Error (RER) (%)			
100	8.47	**8.37 (1.18)**	**8.44 (0.35)**	**7.78 (8.15)**
80	8.53	8.88 (-4.10)	**8.44 (1.055)**	**8.19 (3.99)**
60	8.41	8.83 (-4.99)	8.59 (-2.14)	8.47 (-0.71)
40	8.93	**8.88 (0.56)**	8.97 (-0.45)	**8.91 (0.22)**
20	9.49	**9.08 (4.32)**	9.56 (-0.74)	**8.86 (6.64)**
10	9.66	**9.49 (1.76)**	10.27 (-6.31)	10.27 (-6.31)
5	11.83	**10.37 (12.34)**	**10.56 (10.74)**	**11.24 (4.99)**

These initial results prompt a few lines of further inquiry in addition to the need for improved worst-case robustness already mentioned. Two parameters, kept constant in these experiments, could play an important role in the performance of the different schemes: the augmentation ratio (i.e., how many times we use a transformed version of a base utterance), and the number of distinct transformations injected into the training set. Since each of the 7 transforms described has been applied to the entire training set, both of these parameters are identical in the current experiments. We would like to explore the effect of the first to see if we can establish the point where the utility of augmentation has saturated and/or begun to hurt performance. Similarly, we would like to increase the second by allowing more stochastic variability in the transforms since that could lead to better generalization. Although we purposely omitted any knowledge of the target speaker in our SI experiments to investigate the case of speakers fully unseen at training time, the case where one has access to some data from a target speaker, and no labels, is of interest. In such case, we would like to explore the advantage of DA techniques that instead (or additionally) morph the training set to approximate the target, so as to create data that better match testing conditions. All the these remain the topic of future work.

7. REFERENCES

[1] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of the glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[2] J.H. Jeon and Y. Liu, "Semi-supervised learning for automatic prosodic event detection using co-training algorithm," in *Proc. ACL/IJCNLP*, Singapore, August 2009, vol. I, pp. 540–548.

[3] R. Fernandez and B. Ramabhadran, "Driscriminative training and unsupervised adaptation for labeling prosodic events with limited training data," in *Proc. Interspeech*, Tokyo, Sept. 2010, pp. 1429–1432.

[4] R. Fernandez and B. Ramabhadran, "Exploiting active-learning strategies for annotating prosodic events with limited labeled data," in *Proc. ICASSP*, Prague, May 2011, pp. 2208–2211.

[5] Z. Zhao and X. Ma, "Active learning for the prediction of prosodic phrase boundaries in chinese speech synthesis systems using conditional random fields," in *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, June 2015, pp. 1–5.

[6] A. Gorin, R. Lileikyte, G. Huang, L. Lamel, J-L. Gauvain, and A. Laurent, "Language model data augmentation for keyword spotting in low-resourced training conditions," in *Proc. Interspeech*, San Francisco, 2016, pp. 775–779.

[7] N. Takahashi, M. Gygli, B. Pfister, and L. Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event recognition," in *Proc. Interspeech*, San Francisco, 2016, pp. 2982–2986s.

[8] W Hartmann, T. Ng, R. Hsiao, S. Tsakalidis, and R. Schwartz, "Two-stage data augmentation for low-resourced speech recognition," in *Proc. Interspeech*, San Francisco, 2016, pp. 2378–2382.

[9] N. Jaitly and G.E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML*, Atlanta, 2013.

[10] X. Cui, B. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *Proc. ICASSP*, Florence, 2014, pp. 5582–5586.

[11] Y. Fujita, R. Takashima, T. Homma, and M. Togami, "Data augmentation using multi-input multi-output source separation for deep neural network based acoustic modeling," in *Proc. Interspeech*, San Francisco, 2016, pp. 3818–3822.

[12] A. Rosenberg, "AutoBI - a tool for automatic ToBI annotation.," in *Proc. Interspeech*, Tokyo, 2010, pp. 146–149.

[13] A. Rosenberg, "Modeling intensity contours and the interaction of pitch and intensity to improve automatic prosodic event detection and classification," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Miami, 2012, pp. 376–381.

[14] A. Rosenberg, R. Fernandez, and B. Ramabhadran, "Modeling phrasing and prominence using deep recurrent learning," in *Interspeech*, Dresden, 2015, pp. 3066–3070.

[15] K. Laskowski, J. Edlund, and M. Heldner, "An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems," in *Proc. ICASSP*, Las Vegas, 2008, pp. 5041–5044.

[16] J. Chung, Ç. Gülçehre, K-H. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014.

[17] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM Recurrent Networks," *J. of Machine Learning Research*, vol. 3, pp. 115–143, 2002.

[18] A. Graves, "Generating sequences with recurrent neural networks," *CoRR*, vol. abs/1308.0850, 2013.

[19] E. van den Berg, B. Ramabhadran, and M. Picheny, "Training variance and performance evaluation of neural networks in speech," *CoRR*, vol. abs/1606.04521, 2016.

[20] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," Tech. Rep. ECS-95-001, Boston University, 1996.