FROM TEXT TO PROSODY WITHOUT TOBI

Volker Strom

AT&T Labs Research
180 Park Avenue, Florham Park, NJ 07932-0971
vst@research.att.com

ABSTRACT

A new method for predicting prosodic parameters, i.e. phone durations and F0 targets, from preprocessed text is presented. The prosody model comprises a set of CARTs, which are learned from a large database of labeled speech. This database need not be annotated with Tone and Break Indices (ToBI labels). Instead, a simpler symbolic prosodic description is created by a bootstrapping method. The method had been applied to one Spanish and two German speakers. For the German voices, two listening tests showed a significant preference for the new method over a more traditional approach of prosody prediction, based on hand-crafted rules.

1. INTRODUCTION

Each text-to-speech system first analyzes the input text in order to specify what the speech should sound like, then it generates the output waveform. Text analysis includes part-of-speech (POS) tagging, text normalization, grapheme-to-phoneme conversion, and prosody prediction. Prosody prediction itself often consists of two steps: First, a symbolic description is generated, which indicates the locations of accents and prosodic phrase boundaries (throughout this paper referred to just as "boundaries"). Frequently the symbols are ToBI [1] labels, which are also an abstract description of an F0 (fundamental frequency) contour. From these, the numerical F0 values and phone durations are calculated.

The rational behind this two-step approach is the belief that linguistic features are more strongly correlated with symbolic prosody than with the acoustic realization. This not only makes it easier for a human to write rules which predict prosody, it also makes it easier for a machine to learn these rules from a database.

Unfortunately, ToBI labeling is very slow and expensive [2]. Having several labelers available may speed it up, but it does not address the cost factor. Inter-labeler consistency is another issue [3]. Therefore, a fully automatic procedure is highly desirable. By analogy with automatic phonetic segmentation, which starts out with speaker-independent HMMs and then adapts them to a speaker in an iterative manner, we propose an automatic prosodic labeler, which starts out with speaker-independent (but language-dependent) prosody-predicting rules, and then turns into a classifier, which is iteratively adapted to the acoustic realization of a speaker's prosody. The refined prosodic labels in turn are used to train predictors for F0 targets and phone durations.

2. DATABASE

While in a diphone synthesizer there is only one or a few instances of each diphone which need to be manipulated in order to meet the specifications from the text analysis, in Unit Selection [4] a large database is searched for a sequence of units which meets the specifications best and, at the same time, keeps the joins as smooth as possible. At AT&T, such a database typically consists of several hours of speech per voice. The speech is annotated automatically with words, syllables, phones, and some other features.

The same database is used to train the prosody models. Annotations are enriched with punctuation, POS, and F0. POS tags are generated by the TTS engine. F0 is estimated for each 10 ms frame, and interpolated in unvoiced regions; from the resulting contour, three samples per syllable are taken, at the beginning, middle, and end of the syllable.

We used one American English female speaker, which had ToBI labels for 1477 utterances. They were used to train a prosody recognizer. Its automatically generated labels were used to train the first American English prosody model.

3. CART BUILDING

Our "prosody model" consists of four CARTs (classification and regression trees) [5]: Two of them make binary decisions about where to place accents and boundaries. The other two predict three F0 targets per syllable, and for each phone its z-score, which is the deviation of the phone duration from the mean as a multiple of the standard deviation. The two pairs of CARTs represent symbolic and acoustic prosody prediction respectively. They are made by the free software tool *wagon* [6], applying text-derived features.

For labeling speech with the binary decisions, a different pair of CARTs is used, which applies in addition normalized duration features (see subsection 4.1) as acoustic features.

3.1. Features

A variety of features derived from text are used for prosody prediction. Some refer to words, such as POS, or distance to sentence end. Others refer to syllables, such as stress, or whether the syllable should be accented. For phone duration prediction, additional features refer to phones, for example their phone class or position within the syllable.

Some features are simple, others more complex, such as the "given/new feature". This feature involves lemmatizing the content words and adding them to a "focus stack" [7], which models explicit topic shift; a word is considered "given", if it is already in this stack.

As opposed to more traditional approaches, the binary symbolic prosodic decisions are only two of many features for predicting acoustic prosody: the CART-growing algorithm determines if and when, e.g., the accent feature is considered for predicting the z-score of a specific phone. This way hard decisions on the symbolic level are avoided.

It is known that CART-growing algorithms have problems with capturing dependencies between features. Breiman et. al. [5] suggest combining related features into new features. But trying all possible feature combinations leads to far too many combined features. Providing too many features with most of them correlated often worsens the performance of the resulting CART. A common countermeasure is to wrap CART growing into a feature preselection, but with larger numbers of features this quickly becomes too expensive. The only feasible approach is to offer the feature selection only those relevant combinations suggested in the literature or based on intuition which address the most serious problems.

The final F0 rise in yes-no-questions posed one such problem: Even though the feature set included the punctuation mark, the sentence-initial POS, and whether the sentence contains an "or" (since in alternative questions, the F0 rises at the end of the first alternative, not at sentence end), the CART-growing algorithm was not able to create an appropriate sub tree. This was partly to the sparseness of yes-no-question-final syllables, but even adding copies of did not help: *wagon* needed an explicit binary feature "yes-no question" in order to get question prosody right.

3.2. Quantizing numeric features

While CARTs are an obvious way to deal with categorical features, most CART-growing algorithms cannot really deal with numerical features: Considering all possible splits ($f < c$) is impractical since for each feature f there are up to as many as observed feature values c. Wagon e.g. splits the feature value range in n intervals of equal size. But this kind of quantization may be corrupted by a single outlier. Cluster analysis and quantization up front is the solution in this case.

3.3. F0 target prediction

From the set of F0 vectors (three F0 samples per syllable, see section 2, approximately a dozen clusters are identified by Lloyd's algorithm [8]. The F0 target predictor's task is to predict the cluster index, which in turn is replaced by the centroid vector. The centroid vectors can be seen as prototypes for F0 contours of a syllable. The number of clusters is a trade-off between quantization error and prediction accuracy. It is also important to cover rare but important cases, e.g. the final rise in yes-no questions. This can be done by equalizing the training data.

4. ITERATIVE CART GROWING

The basic idea in iterative CART growing is to alternate between prosody prediction from text and prosody recognition from text plus speech. To that end, it is a special case of the Expectation Maximization algorithm [9].

Initial accent and boundary labels are obtained by simple rules: Each longer pause is considered a boundary, as well as each sentence boundary (most of which coincide with a pause). ToBI hand labels for a large corpus of one female American English speaker suggest that boundaries and pauses are highly correlated.

As far as accents are concerned, we would like to initialize the iteration with a speaker-independent accent recognizer, as it is the case with the simple boundary recognizer. Acoustic cues for accents are less strong, and some are similar to cues for boundaries (see subsection 4.1). For now, initial accent labels are made applying a simple rule on text-derived features only. Care must be taken that after the first iteration the resulting CART does not reflect just this rule, i.e. does not look at acoustic features at all. This can be achieved by switching between "sufficiently orthogonal" feature sets.

Once there is a prosody model for a speaker of the same language, one can use it to obtain initial labels. A more general and speaker-independent prosody model can be achieved by further pruning the corresponding CARTs.

The predicted accent and boundary labels are added to the feature vectors. From this data, the first CARTs predicting durations and F0 targets are made. Often they already produce better sounding prosody than hand-crafted rules, probably because they are inherently speaker-adaptive.

4.1. Normalized durations

Once CARTs exist that predict durations and F0 from text, these models can be used to refine the accent and boundary labels by looking not only at textual features, but also at acoustic features. In the second step, more acoustic information is available than just the presence or absence of a pause. The second most important acoustic feature is the relative syllable duration. Accented syllables as well as phrase-final syllables are lengthened. Thus, accent and boundary models must be refined simultaneously. The amount of lengthening is determined by the ratio of actual and predicted duration.

In the same manner, actual and predicted duration of a whole prosodic phrase can be compared, which allows for some degree of speaking rate normalization. A new CART is grown which predicts these normalized durations. To that end, improving the duration model itself is an iterative process.

4.2. Prosodic labeling

Pause durations and syllable durations, obtained from the phonetic segmentation and normalized with respect to speaking rate and intrinsic duration are added to the textual features. Eleven further features are extracted from the speech signal: Three median-smoothed energy bands derived from the log. short time FFT make the energy features. The interpolated F0 is decomposed into 3 components with band pass filters. The F0, their 3 components, and their time derivatives of them make eight features, that describe the F0 contour locally and globally [10].

A classifier then is trained to recognize the accent and boundary labels predicted in the previous step. With a mixed set of features, the problem is that CARTs cannot really handle numerical features (see subsection 3.2) and numerical classifiers cannot deal with categorical features (unless it is a binary feature encoded as 0/1, plus a little noise in order to avoid numerical problems). Therefore a hierarchic classifier was chosen: The CARTs predicting accents and boundaries from text features can not only output the class having the highest posterior probability, but also this probability itself. The two probabilities are then added to the acoustic features as "linguistic features", and a numeric classifier is applied.

With the hand-labeled data for the female American English speaker it was found that an n-nearest-neighbor classifier works best. Its accent and boundary labels are correct (with hand labels as reference) for 88.5% and 96.7% of all syllables respectively, which is close to the inter-labeler consistency.

The machine labels are then fed into the next iteration step, growing prosody-predicting CARTs. With the speakers dealt with so far, the prosodic labels created this way stabilized quickly during the iteration, so that two iterations seem to be enough at this point. However, more research is needed with other speakers.

Even with the optimal set of features, predicted and recognized prosody labels will never fully converge, because there are many ways of saying something correctly. But the goal here is to optimize prosody *prediction*.

For example, our German male speaker often paused at places where one would normally not pause. This resulted in initial boundary labels which were too difficult to predict from text. Fortunately, there was already a reasonable CART for the German female speaker, which was substituted for the first iteration.

There was some trouble with the female speaker, too: When examining her yes-no questions, a few ended with a falling F0. Some of them were errors of the F0 extraction, but in others, the F0 actually fell. One may argue that these are not really yes-no questions. For some cases the corresponding feature could be improved (e.g. wh-questions which start with an excuse), the remaining ones would require the TTS engine to "knowing what it is saying". For now, it is best to remove examples from the training data that are too difficult.

5. PERCEPTUAL EVALUATION

The ultimate goal in TTS is to improve the overall quality. Evaluating individual components, e.g. looking at the RMSE of the phone duration predictor, may give clues about how to improve it, but due to interaction with unit selection, decreasing the RMSE does not necessarily improve the overall quality: When predicting "extreme" durations, there are fewer units in the database and a unit selected for duration may fit badly in some other respect. Some prosody setting may improve quality on the segmental level, even at the cost of higher RMSEs. It is still a challenge to find measures for acoustic distance which are perceptually more relevant. For now, TTS quality is still assessed best by human listeners.

Twelve adult native speakers of German participated as volunteers in a web-based listening evaluation of several different versions of German TTS. There were two parts to the web-based evaluation experiment, the first part was an A/B Paired Comparison Test, and the second, a subjective rating test.

5.1. Paired comparison listening test

Two versions of prosody were compared: **ManPro**, the hand-crafted prosody rule set, and **DataPro**, the rule set generated automatically from data. There is one DataPro for each individual speaker, while one ManPro per language, which is adapted to the speaker's pitch range.

There are two voices for the AT&T German TTS system. The input text for 19 test utterances was selected completely independently of and prior to synthesis by any TTS system tested. The test utterances included: 10 interactive prompts and 3 German news paragraphs, which were edited into 9 separate test utterances for the purposes of paired comparison tests.

Three of the 12 listeners were familiar with German TTS and nine were unfamiliar. All were employees of AT&T. All subjects were blind as to the identity of the synthesis system associated with each utterance. After listening to the two versions of an utterance, they were instructed to click on the icon representing the utterance they believe sounded best. Participants were told that they may listen to the utterances in any order, and as many times as they liked.

DataPro was favored over I prosody by nearly a 2:1 ratio. The mean number of utterances preferred (and the corresponding percentage of preferred from among 19 test trials) is shown by voice for each of the two TTS prosody modules in table 1. Across voices, a preference for DataPro over ManPro was shown by 11 of the 12 listeners, and no listener had a preference for ManPro.

Results of separate one-sample t-tests (two-tailed) for each voice (with the null hypothesized value of mu = 9.5, which is half of the 19 test trials per voice) indicate a significant preference for DataPro over ManPro (Klara: $t = 3.4039$, $df = 11$, $p = 0.0059$; Reiner: $t = 5.3759$, $df = 11$, $p = 0.0002$). Similarly, the 65.1% preference for DataPro was significant (with mu = 19, which is half of the 38 total test trials) when preferences were summed across both voices ($t = 5.8335$, $df = 11$, $p = 0.0001$).

VOICE	DataPro Pref.	ManPro Pref.	95% Conf.Int.
Klara	12.4 (65.4%)	6.6 (34.6%)	10.53 - 14.30
Reiner	12.3 (64.9%)	6.7 (35.1%)	11.17 - 13.49
Pooled	24.8 (65.1%)	13.3 (34.9%)	22.58 - 26.92

Table 1. *Paired comparison listening test, (19 test utterances).*

6. SUBJECTIVE RATINGS OF GERMAN TTS PARAGRAPHS

Only the three complete German news paragraphs were used as text input for the ratings portion of the evaluation. Five test utterances were synthesized for each test paragraph: (1) Klara voice with DataPro, (2) Reiner voice with DataPro, (3) Klara voice with ManPro, (4) Reiner voice with ManPro, (5) female German voice from the highest quality commercial competition available for comparison.

Participants were told, as before, to listen to an utterance by clicking on its icon. Listeners were blind as to the identity of the synthesis system that generated each utterance. After listening to an utterance, they were instructed to click on the icon corresponding to the rating – on a scale from 1 (Bad) to 5 (Excellent) – that they believed to best represent the speech quality of the utterance. Again, they could listen to the utterances in any order, and as many times as they liked, and they were encouraged to use headphones.

For both AT&T TTS voices, DataPro was rated 0.35 MOS higher on average than ManPro. A repeated measures ANOVA was performed, with TTS system (5) and paragraph (3) the within-subject factors in the fully factorial design. Each prosody version for each voice of AT&T TTS was rated significantly higher than competitor's TTS. Ratings of DataPro were consistently higher than ManPro ratings for both voices, but only for the Reiner voice did the size of the difference reach statistical significance. Comparing ratings for the two AT&T German TTS voices, Klara's ratings were significantly higher than Reiner's ratings, regardless of the prosody version used. Table 2 below lists the mean opinion scores (MOS), standard errors, and lower and upper bounds of the 95% confidence intervals for each TTS condition tested.

Voice	Prosody	MOS	SE	95%Conf.Int.
Klara	DataPro	3.556	0.155	3.214 - 3.897
Klara	ManPro	3.250	0.179	2.855 - 3.645
Reiner	DataPro	3.194	0.145	2.876 - 3.513
Reiner	ManPro	2.806	0.166	2.439 - 3.172
female	Compet.	1.972	0.186	1.564 - 2.381

Table 2. *Subjective Ratings of three news paragraphs.*

7. SUMMARY

A method has been described that learns text-to-prosody from speech data. All annotations are made fully automatically from text. The prosodic annotations are created by a bootstrapping method and indicate just the locations of accents and boundaries. During synthesis, when predicting durations and F0 from features derived from text, predicted accents and boundaries serve as additional binary features only.

The method adapts itself to each individual speaker. Prosody predictors based on hand-crafted rules typically allow adaption to only a few parameters for to a specific speaker, such as phone duration means and standard deviations, and F0 topline and baseline; they are not able to capture more prosodic characteristics of the speaker.

The prosody predictors are created in an iterative method by alternating prosodic labeling and prosody prediction. For the two German voices, two iterations and modest manual interference resulted in CARTs that predict prosody significantly better than a hand-crafted rule set, as two listening tests showed.

Applying the method will always require some manual work, such as adapting some linguistic features and creating a fairly speaker-independent prosody model for that language. Adding just a new speaker is a far more easy task. Future work will focus on fully automating the process.

8. ACKNOWLEDGMENTS

Thanks to Ann Syrdal for conducting and evaluating the listening tests.

9. REFERENCES

[1] K. Silverman, M. Beckman, J. Pitrelli, M. Osterndorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, “Tobi: A standard for labeling English prosody,” in *Proc. Int. Conf. on Spoken Language Processing*, 1992, pp. 867–870.

[2] A. Syrdal and J. Hirschberg, “Automatic ToBI prediction and alignment to speed manual labeling of prosody,” *Speech Communication, Special Issue on Speech Annotation and Corpus Tools*, , no. 33, pp. 135–151, 2001.

[3] A. Syrdal, “Inter-transcriber reliability of ToBI prosodic labeling,” in *Proc. Int. Conf. on Spoken Language Processing*, Beijing, 2000.

[4] A. Conkie, J. Schroeter, Y. Styliano, and A. Syrdal, “The AT&T Next-Gen TTT System,” in *Proc. Joint Meeting of ASA, EAA and DEGA*, 1999.

[5] L. Breiman, J. Friedman, R. Olshen, and C. Stone, “Classification and Regression Trees,” Boca Raton, 1984.

[6] A. Black, R. Caley, S. King, and P. Taylor, “CSTR software,” http://www.cstr.ed.ac.uk/software.

[7] J. Hirschberg, “Pitch accent in context: predicting intonational prominence from context,” in *Artificial Intelligence*, 1993, pp. 305–340.

[8] S. P. Lloyd, “Least squares quantization in PCM,” in *IEEE Trans. on Inf. Theory*, 1982, vol. 28, pp. 129–137.

[9] A.P. Dempster, N.M. Laird, and D.B. Rubin, “Maximum likelihood from incomplete data via the EM Algorithm,” *Journal of the Royal Statistical Society*, vol. 39, pp. 1–38, 1977.

[10] V. Strom, “Detection of accents, phrase boundaries and sentence modality in German with prosodic features,” in *Proc. European Conf. on Speech Communication and Technology*, Madrid, 1995, vol. 3, pp. 2039–2041.