Using Automatic Stress Extraction from Audio for Improved Prosody Modelling in Speech Synthesis

György Szaszák¹, András Beke², Gábor Olaszy³ and Bálint Pál Tóth¹

¹Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics
²Research Institute for Linguistics, Hungarian Academy of Sciences
³Research Institute for Linguistics, Budapest University of Technology and Economics

Abstract

Generating proper and natural sounding prosody is one of the key interests of today’s speech synthesis research. An important factor in this effort is the availability of a precisely labelled speech corpus with adequate prosodic stress marking. Obtaining such a labelling constitutes a huge effort, whereas inter-annotator agreement scores are usually found far below 100%. Stress marking based on phonetic transcription is an alternative, but yields even poorer quality than human annotation. Applying an automatic labelling may help overcoming these difficulties. The current paper presents an automatic approach for stress detection based purely on audio, which is used to derive an automatic, layered labelling of stress events and link them to syllables. For proof of concept, a speech corpus was extended by the output of the stress detection algorithm and a HMM-TTS system was trained with the extended corpus. Results are compared to a baseline system, trained on the same database, but with stress marking obtained from textual transcriptions after applying a set of linguistic rules. The evaluation includes CMOS tests and the analysis of the decision trees. Results show an overall improvement in prosodic properties of the synthesized speech. Subjective ratings reveal a voice perceived as more natural.

Index Terms: prosody analysis, stress detection, speech synthesis, prosody generation, automatic prosody labelling

1. Introduction

During the preparation of a speech corpus for Text-To-Speech (TTS) purposes, a basic step is the precise labelling of prosodic stress. This can be done manually or automatically. Manual stress annotation is relatively precise, however, it is very time consuming and introduces subjectivity. For example, inter-annotator agreement scores using ToBI annotation are usually found between 70-80% for pitch-accents in English [1]. Our personal experience shows that syntax influences human annotation highly, even if done by high-qualified experts. Other results suggest that if stress is syntactically predictable, its marking may be missed by prosodic features [2].

An alternative way of stress labelling is automatic, and based on the transcription of the speech material. This approach can also lead to a corpus suitable for good quality synthesis, however, it also suffers from errors resulting from the mismatch between syntactic and prosodic marking of stress. If stress marking is done in a rule-based manner as is often the case, further difficulty evolves when stress prediction fails, due to the well known lack of generalization capabilities of the rule-based approach. However, this latter problem can be cured, if speech corpora for TTS are well planned and read carefully, hence utterances requiring special attention can be handled by exceptions.

Given these difficulties of stress labelling, and the fact that even human labelling is somewhat ambiguous (inter-annotator agreements up to 80%), an approach predicting stress based on audio could be used to obtain stress labelling. Acoustic-prosodic features have already been used for automatic labelling of TTS inventories [3], [4] often relying on ToBI [5], [6]. These approaches rely on syllable level data-driven classification, whether stress (accent) is present or not on a given syllable. Although most of TTS systems require a syllable level annotation of stress and human perception is also capable of assigning the stress to a syllable, our assumption is that stress, being a prosodic event is better treated as a supra-segmental phenomenon. This approach is also sustained by speech production models [7], where final prosody is complex and is influenced on multiple levels in the top-down like speech production process.

For this reason, a novel top-down approach is proposed in this paper, instead of handling the problem directly and exclusively on the syllable level [8]. The proposed approach relies on the automatic detection of phonological phrases. Given that phonological phrases are units characterized by own stress and a more or less complete intonation contour, an exact recovery of the phonological phrase structure can be made equivalent with stress detection. See Section 3 for more details. Another advantage of choosing a unit much higher than a syllable is that tendencies in intonation become also traceable and detectable, within the same framework, which allows for a uniform and very simple approach suitable for complex prosodic analysis.

This paper is organized as follows: first, the frameworks for automatic stress detection based on transcripts and on the audio signal are presented. An experiment is run to compare these two approaches. Hereafter, results are presented and evaluated, including decision tree analysis of the TTS system and subjective listening tests. Finally, conclusions are drawn.

2. Stress Generation from Transcripts

Stress detection based on the transcript takes place relying on rules applied after morphological and syntactic analysis, performed using a dependency grammar [9]. Stress marking takes place relying on the output of this analysis. The complex description of the text based stress marking algorithm can be found in [10], here just the basic operation is illustrated. Four levels of stress are distinguished: (i) very strong stress, assigned based on an exception list, typically covering contrastive negation; (ii) strong stress, also assigned based on an exception list;
(iii) medium stress, this level of stress is assigned by rules; (iv) finally, the fourth level of stress is the neutral (unstressed) form.

The most important rules involved for medium stress prediction include: (i) sentence initial words are medium stressed – exceptions may apply; (ii) all words following an article or the conjunction word “és” (“and”) are stressed; (iii) all words following a comma are stressed – exceptions may apply based on a list; (iv) the last word of a sentence is never stressed; (v) articles and words following a strongly stressed word cannot be stressed.

In current work, very strong and strong stress are merged, which means a 3 level rating for syllables: strongly stressed, stressed and neutral. Hereafter, this labelling based on phonetic transcriptions is referred to as TBSM (Text Based Stress Modelling).

3. Automatic Stress Detection from Audio

This section describes automatic stress detection from the speech signal, using a simple and effective phonological phrase (PP) modelling approach. PPs can be found under intonational phrases (IP) in the prosodic hierarchy. Intonational phrases build up utterances [11].

PPs constitute a prosodic unit, characterized by an own stress and some preceding/following intonation contour. Keeping in mind the limitations of the syntax/phonology interface, they can be thought of corresponding often to a group of words, a clitic group in the speech transcription.

Given that PPs have a specific acoustic-prosodic contour, they can be classified and hence modelled separately. Relying on a data-driven machine learning approach, PP models can be trained and used for classification. We choose the Hidden Markov Model / Gaussian Mixture Model (HMM/GMM) framework for PP modelling and use a Viterbi alignment for the utterances requiring stress detection. This choice is motivated primarily by the excellent time warping capabilities of HMMs. The overall approach is documented in [8], in this paper a brief review is provided for the sake of comprehension.

3.1. Fundamentals of PP Modelling

Our experiments are planned for the Hungarian language, where 7 different PP types were created in [8], as listed in Table 1.

<table>
<thead>
<tr>
<th>Label</th>
<th>Stress</th>
<th>Contour</th>
</tr>
</thead>
<tbody>
<tr>
<td>co</td>
<td>strong</td>
<td>Clause onset then descending</td>
</tr>
<tr>
<td>ss</td>
<td>strong</td>
<td>Stress then descending</td>
</tr>
<tr>
<td>ms</td>
<td>medium</td>
<td>Stress then descending</td>
</tr>
<tr>
<td>ce</td>
<td>medium</td>
<td>Stress then low ending</td>
</tr>
<tr>
<td>cr</td>
<td>medium</td>
<td>Stress then ascending</td>
</tr>
<tr>
<td>ls</td>
<td>neutral</td>
<td>No or negative stress + descending</td>
</tr>
<tr>
<td>sil</td>
<td>neutral</td>
<td>Silence</td>
</tr>
</tbody>
</table>

The distinction between PPs consists of two components: (i) the strength of stress the PP carries and (ii) its characteristic intonation contour (see Table 1). Descending vs. falling and ascending vs. rising F0 contours are not separated for this experiment. The theoretical prototype of a PP in Hungarian shows a smart rise of F0 at the stressed syllable, followed by a gradually descending intonation contour (ms). As Hungarian is a fixed-stress language (stress, if present, is bound to the first syllable of a word), location of the stress within the PP is not a distinctive feature.

3.2. Acoustic Features

Acoustic-prosodic features used for PP modelling include fundamental frequency (F0), wide-band energy (E) and optionally, syllable duration. Syllable duration is not used for Hungarian, as it was not found to be a distinctive cue regarding stress [8].

F0 extraction is performed with the compute-kaldi-pitch tool of the Kaldi ASR toolkit, which provides an overall continuous F0 contour, defined for unvoiced frames as well [12]. Compared to other pitch trackers, this tool has very positive properties: constraining allowed F0 values throughout the whole utterance in a Viterbi decoding like approach and effectively limiting sudden changes it provides a stable F0 estimate, with consequent behaviour over time.

For energy computation a standard integrating is applied with a window span of 150 ms. Frame rate is 10 ms. First and second order deltas are added to both F0 and E streams.

3.3. HMM/GMM for PPs

PPs are modelled by left-to-right HMMs with 11 states. GMMs associated to emitting states usually have 1..4 mixture components and accept 6 dimensional observations.

The training of PP models is performed in a supervised manner on a sub-corpus of the Hungarian BABEL corpus, labelled for phonological phrases. 300 sentences from 24 speakers are used.

When used for stress detection, PP models are aligned to the utterance with a looped grammar allowing each kind of PP to occur with equal probability. This provides a global phonological phrasing of the utterance(s). Stress detection is based on the PP boundaries. For the stress bound Hungarian, stress is detected on the syllable right after the PP boundary (first syllable). Stress modelling based on audio is hereafter referred to as ABSM (Audio Based Stress Modelling).

We believe this framework is adaptable to other languages, with a language dependent component: the PP models. In stress unbound languages, a further step is necessary for the stress detection based on phonological phrasing: using a state level alignment, the state capturing the stress should be mapped to the underlying syllable. If a syllabification is available with time stamps, this is a trivial task.

Please note that with the PP alignment approach, a continuous tracking of prosodic structure over the whole speech signal is implemented instead of looking for discrete markers, indices of breaks or tones. This provides a soft and flexible framework, which is also believed to be closer to human perception processes.

3.4. Performance Evaluation of the PP Classification

Using a 10-fold cross-validation, the PP alignment algorithm is evaluated on the BABEL corpus: a PP alignment is generated which is then compared to the reference obtained by hand-labelling. Detection is regarded to be correct if the boundary is detected within the $TOL=100$ ms vicinity of the reference. The following performance indicators are computed: (i) recall ($RCL$) and (ii) precision ($PRC$) of the PP boundary detection and (iii) the average time deviation ($ATD$) between the detected and the reference PP boundary. Fig. 1 shows detection performance depending on $ATD$ for operating points where
Figure 1: Precision and recall rates in operating points defined by \( PRC = RCL \) depending on the tolerance value TOL.

Precision and recall are equal \( (PRC = RCL) \). In the operating point defined by \( PRC = RCL = 71.0\% \) \( ATD \) is 31.9 ms. If we allow a tolerance interval of \( TOL = 200 \text{ ms} \), still comparable to typical syllable length, we obtain \( PRC = RCL = 84.8\% \) by \( ATD = 54.3 \text{ ms} \).

4. Experimental Setup

4.1. Automatic Labelling of the TTS Corpus

The TTS corpus is a parallel, Precisely Labelled Hungarian corpus (PLHD) containing 1984 sentences uttered by 14 speakers [13]. Precise labelling covers manual phonetic transcriptions and phone level segmentation, but does not contain hand labelling of stress in the currently available version.

The TTS corpus is labelled for stress automatically in two different ways, keeping the obtained labellings separately. Labelling based on text is performed as described in section 2. Labelling based on audio is performed as follows: after prosodic feature extraction, an automatic PP alignment is run for all utterances. As Hungarian is fixed stressed on the first syllable, each syllable following a detected PP boundary is marked for stress. Syllabification is obtained automatically based on text, syllable boundaries are derived from phone level segmentation. Stress marking occurs depending on the strength of stress: two types of stress are distinguished, a strong and a medium one (see Table 1). All other syllables are kept neutral.

4.2. Proof of concept TTS systems

Two HMM-based TTS systems were trained to compare the effects of TBSM and ABSM on the prosody of the synthesized speech. With a modified version of the HTS toolkit [14] a female voice (48 kHz, 16 bits) was trained from the PLHD TTS corpus. The training database contained the same 1984 sentences and a manually verified phonetic transcription for both scenarios. The questions used for building the decision trees were the same. MDL-based (Minimum Description Length) decision tree pruning was used [15]. The only difference between the two systems was the stress labels in the training database. In one case stress was marked in the training database based on the phonetic transcription. This system will be referred to as TBSM-TTS. In the other case stress labels were determined from pure audio. This second case will be denoted by ABSM-TTS. During speech synthesis, phonetic transcription based stress label prediction was used in both cases.

5. Results

The difference between the F0 contour of the synthesized speech and underlying syllable-level stress patterns is illustrated in Fig. 2 for the TBSM and ABSM case, respectively.

In order to provide a deep overview of the effect of using ABSM instead of TBSM, results are presented from two points of view, an objective, and a subjective one. For objective evaluation, the resulted decision trees are analysed and compared in order to see how prosody related questions are exploited in the two systems. The subjective evaluation involves listening tests where human volunteers should rate the perceived difference, focussing on prosody, of TBSM-TTS and ABSM-TTS case.

5.1. Analysis of the Decision Trees

The structure of the decision trees that were created automatically during the training of the TBSM- and ABSM-TTS systems is investigated to explore the possible enhancements of the proposed method. In the decision trees, stress related features are examined covering questions like (i) is the previous/current/next syllable strongly stressed/stressed/neutral, (ii) what is the number of syllables from/to the previous/next stressed syllable. The analysis included the decision trees of fundamental frequency \( (F0) \) and phone durations. The trees of spectral components contained hardly any stress related feature, as expected, thus they were excluded from this analysis.

Table 2: Total number of nodes and stress related nodes in \( F0 \) and duration decision trees of TBSM- and ABSM-TTS.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Tree nodes</th>
<th>( F0 )</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBSM</td>
<td>Total</td>
<td>Stress</td>
<td>4521</td>
</tr>
<tr>
<td>ABSM</td>
<td>Total</td>
<td>Stress</td>
<td>4653</td>
</tr>
</tbody>
</table>

As Table 2 shows, training with ABSM resulted in larger decision trees (the sizes of \( F0 \) and duration decision trees are increased by 2.9% and 4.1%, respectively), which suggests better, more detailed modelling capabilities. Not only the deci-
The results are shown in Fig. 4. Altogether 400 sample pairs were rated (20 test subjects ≥ 20 sample pairs). 51 votes out of 400 (12.75% of ratings) found the ABSM examples much more natural than the TBSM version, whereas 6.75% of the ratings judged the TBSM examples much more natural than ABSM examples. 27.75% of the ABSM examples were perceived more natural than their TBSM counterpart, whereas 25.75% of the TBSM utterances were found more natural. The results were investigated by Student-t test (α = 0.05) and showed a significant improvement. (A probability value of less than alpha was considered statistically significant.)

During the subjective tests we noted that subjective ratings can hardly be focussed on prosody, but are mostly influenced by overall speech quality. Hence, it is difficult to evaluate changes in prosody separately. Even when F0 plots of ABSM and TBSM systems showed high, grammatically correct difference, listeners often chose “equal” quality. This may be due to the unchanged voice quality regarding spectral-segmental speech properties. But despite this difficulty, ABSM-TTS system was overall perceived significantly more natural than TBSM-TTS.

### 6. Conclusions

This paper presented an automatic method for stress labelling based on audio. Phonological phrases were modelled based on prosodic-acoustic features and used in an alignment mode to perform stress detection for a TTS corpus. Precision and recall of the PP alignment ranged between 71.0% and 84.8%, depending on the tolerance interval (ranging 100-250 ms) applied when comparing to the reference boundary. The approach was successfully used in training a HMM-TTS system and significantly outperformed the baseline system trained on stress labels generated based on phonetic transcription. Performance improvement was analysed in the decision trees, showing more stress-related questions ranked higher with ABSM approach. Subjective listening tests also confirmed a significant improvement.

The amount of hand-labelled data required to train the audio based stress detector was 6 times smaller than the TTS corpus for a single speaker. The hand labelling of the stress detector’s training corpus consists in PP segmentation, which is simpler than a complex prosodic annotation such as ToBI.

The algorithm is suitable to be used for any TTS technology: unit selection or data-driven (parametric). The ABSM approach has a language dependent module modelling PPs, but given PPs are almost universally found in all languages, the technology has good chances to be relatively easily adaptable – however, this facility needs further experimental confirmation.

### 7. Acknowledgements

György Szaszká would like to thank the support of the Hungarian Scientific Research Fund (OTKA) under contract ID PD-112598, Automatic Phonological Phrase and Prosodic Event Detection for the Extraction of Syntactic and Semantic/Pragmatic Information from Speech. Bálint Pál Tóth assisted in fundamental research in the frame of TAMOP 4.2.4. A/1-11-1-2012-0001 National Excellence Program Elaborating and operating an inland student and researcher personal support system, realized with personal support. The project was subsidized by the European Union and co-financed by the European Social Fund. The authors express their gratitude for the support of the Swiss National Science Foundation via the joint research project (SCOPES scheme) “SP2: SCOPES project on speech prosody” (SNSF N° 17320-152495/1).
8. References


