Initial Experiments on Automatic Correction of Prosodic Annotation of Large Speech Corpora*

Zdeněk Hanzlíček and Martin Grůber

NTIS - New Technology for the Information Society,
Faculty of Applied Sciences, University of West Bohemia,
Univerzitní 22, 306 14 Plzeň, Czech Republic
{zhanzlic,gruber}@ntis.zcu.cz
http://www.ntis.zcu.cz/en

Abstract. Most modern speech synthesis systems utilize large speech corpora to learn new voices. These speech corpora usually contain several hours of speech spoken by talented speakers who are able to record such an amount of speech data in a sufficient quality. An appropriate phonetic and prosodic annotation of the recorded utterances is necessary for a high quality of synthesized speech. For many languages, the pitch shape within the last prosodic word of a phrase is characteristic for particular types of sentences and phrase structure of compound/complex sentences. However in the real data, this formal convention can be breached and a different pitch shape than expected can be present. This can be a source of prosody inconsistency in synthesized speech. This article presents some experiments on automatic detection of prosodic mismatch in recorded utterances. A simple classifier based on GMM was proposed for this task. Experiments were performed on 5 large speech corpora. The classification results were successfully verified by listening tests.

Keywords: speech corpora, prosodic annotation, prosodeme.

1 Introduction

Most modern speech synthesis systems [1,2] utilize large speech corpora to learn new voices. These speech corpora usually contain several hours of speech spoken by talented speakers who are able to record such an amount of speech data in a sufficient quality. An appropriate phonetic and prosodic annotation of the recorded utterances is necessary for a high quality of synthesized speech [3]. Generally, the knowledge of presence of various prosodic events in speech data and their detailed description can be useful for many other applications as well.

In connection with using the large speech corpora, the automatic phonetic and prosodic annotation of speech [4,5] became an important task. This article presents some initial experiments on automatic detection of prosodic mismatch in recorded utterances.

* This work was supported by the Technology Agency of the Czech Republic, project No. TA01011264 and by the European Regional Development Fund (ERDF), project "New Technologies for Information Society" (NTIS), European Centre of Excellence, ED1.1.00/02.0090.

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 481–488, 2014.

For many languages, the pitch shape within the last prosodic word of a phrase (corresponding to functionally involved prosodeme[1]) is characteristic for particular types of sentences and for the phrase structure of compound/complex sentences. However, in the real speech data, this formal convention can be breached and a different type of prosodeme than expected can be present. Using a speech corpus with bad prosodeme labels can be source of prosody inconsistency in synthesized speech. Prosodemes whose type does not correspond to the given sentence structure should be revealed and corrected or removed from the corpus. This should improve the overall quality of resulting synthetic speech.

This paper is organized as follows, Section 2 explains the prosody model used in this work. Procedure for prosodeme classification is proposed in Section 3. Section 4 describes performed experiments and their results. Finally, Section 5 concludes this paper and outlines the future work.

2 Prosody Model and Prosodemes

Within this paper, the formal prosody model proposed by Romportl [6] is used. According to this model, an utterance can be divided into prosodic clauses separated by short pauses. Each prosodic clause includes one or more prosodic phrases, which contain certain continuous intonation scheme. A prosodic phrase consists of two prosodemes: null prosodeme and functionally involved prosodeme which is usually related to the last prosodic word in the phrase.

For the Czech language[2], the following basic classes of functionally involved prosodemes are distinguished (for detailed prosodme categorization see [6]):

P1 – prosodemes terminating satisfactorily (specific for declarative sentences)
P2 – prosodemes terminating unsatisfactorily (specific for questions)
P3 – prosodemes non-terminating (specific for non-terminal phrases in compound/complex sentences)

This paper is focused on compound/complex sentences. We assume that the last phrase in these sentences ends with prosodeme P1 and all previous phrases end with prosodeme P3. In the case of neutral speech (no emphasis, expression etc.), prosodemes P1-1 and P3-1 are expected. Typical examples of prosodemes P1-1 and P3-1 are depicted on Figures 1 and 2.

A typical feature for prosodeme P1-1 is a pitch decrease within its last syllable. For prosodeme P3-1, a pitch increase within the last syllable is specific. In some cases, the pitch increase/decrease can be realized as a value contrast between pitch of last and previous syllable.

Beside the pitch shape, spectral, duration and energy features can be characteristic for particular prosodemes. However, their impact seems to be not so relevant for prosody perception or the dependence is more complex.

In real speech data, a different prosodeme than expected could be present. This problem appears even in utterances spoken by a professional speaker. A typical example

[1] Prosodemes are described in Section 2.

[2] A different/modified set of prosodemes can be specific for other languages.

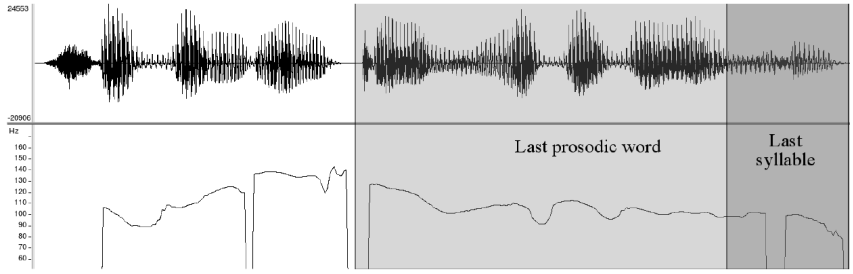

Fig. 1. Example of a phrase terminated with prosodeme P1-1 (waveform and pitch)

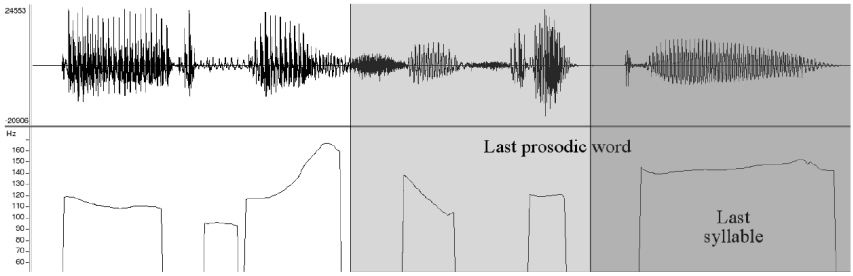

Fig. 2. Example of a phrase terminated with prosodeme P3-1 (waveform and pitch)

is the compound sentence that can be split into several independent sentences. Within the compound sentence, all phrases (except the last one) should be terminated with the prosodeme P3-1. However, the independent sentences are naturally terminated by the prosodeme P1-1.

Badly annotated corpus can be a source of various troubles in some applications. In speech synthesis (specifically, in unit selection method), prosodeme labels are used for selecting sequence of optimal speech units for building resulting speech [7]. Using units from an inappropriate prosodeme or mixing units from different types of prosodemes can cause a decrease in the overall speech quality – prosody of synthesized speech does not correspond to the type or the structure of the sentence, some unnatural pitch fluctuation can occur, etc.

3 Prosodeme Classification

Gaussian mixture models (GMMs) are widely used in various speech classification tasks, such as speaker identification [8], emotion recognition [9], etc. Since the usage of GMMs is straightforward and the performance is usually satisfactory, we decided to use them in our experiments as a baseline.

First, each prosodic phrase is represented by a feature vector F and a default prosodeme type P_X. Pitch is extracted from audio files by using the RAPT algorithm [10] implemented in the SPTK toolkit [11]. The feature vector is computed from the extracted pitch as follows

1. The pitch within the whole phrase is normalized to zero mean.
2. The average values of pitch within the last and the last but one syllabic core are calculated: $\bar{f_1}$ and $\bar{f_2}$, respectively. Since both values are calculated from normalized pitch, they express the emphasis within the last two syllables.
3. The slope of pitch df_1 within the last syllabic core is determined by linear regression.
4. The final feature vector F is composed as

$$F = \left[\bar{f_1}, (\bar{f_1} - \bar{f_2}), df_1\right]^T$$

For each prosodeme P_X, a simple Gaussian mixture model $\mathcal{G}_{P_X}(F)$ is trained. Moreover, the weigh $\mathcal{W}_{P_X}$ for particular models is given as a relative number of corresponding prosodeme in the training data.

$$\mathcal{W}_{P_X} = \frac{\text{number of prosodeme } P_X}{\text{number of all prosodemes}}$$

Classification decision is done by

$$P_Y = \arg\max_{\{P_X\}}\left[\mathcal{W}_{P_X}\mathcal{G}_{P_X}(F)\right]$$

Since all the values in the feature vector F are calculated from the normalized pitch, they seems to be (partly) speaker-independent. Thus, it would be possible to train one speaker-independent set of classifiers; however, to capture the speaker specific features more precisely, individual classifiers were used for particular speakers in our initial experiments.

The training and the classification are performed on whole speech corpora. This approach is based on the assumption that most prosodemes are correct and the minority of incorrect prosodemes should not influence the training of the classifiers since as outliers they are not taken into account. In the case of large amount of incorrect prosodemes, the classifiers would be probably poorly trained and would be inapplicable.

4 Experiments and Results

4.1 Experimental Data

For our experiments, we used 5 large speech corpora recorded for the purposes of speech synthesis [12]: 3 male voices (denoted as M_{AJ}, M_{JS}, M_{TJ}) and 2 female voices (denoted as F_{MR}, F_{KI}). Each corpus contains about 10,000 utterances[3]. With the exception of M_{TJ}, all corpora contain the same sentences[4]; corpus M_{TJ} was partly different.

Since a proper phonetic segmentation [4] (including pauses) was available for all corpora, splitting particular utterances into prosodic clauses was straightforward.

[3] Particular corpora contained larger number of utterances, but only declarative sentences were selected for our experiments. Thus, the accurate number of utterances selected from particular corpora corresponds to the number of prosodemes P1-1.

[4] The numbers of utterances slightly differ because some defective utterances were discarded.

Table 1. Number of prosodemes in particular corpora

prosodeme	M_{AJ}	M_{JS}	M_{TJ}	F_{MR}	F_{KI}
P1-1	10,001	9,896	9,896	9,897	9,878
P3-1	13,051	10,545	17,479	8,581	3,562

However, dividing clauses into phrases is a more complex task because a sophisticated text analysis is necessary – a simple detection of conjunctions and punctuation marks is not sufficient. From that reason, we decided to perform our initial experiments only on the last phrase in each clause where the prosodeme occurrence was guaranteed.

A simple prosodic profiles of particular corpora are presented in Table. 1. Different values illustrate various speaking styles of particular speakers. Since only last phrases in particular prosodic clauses were taken into account, the number of prosodeme P3-1 corresponds to the number of pauses in speech.

4.2 Classification Results

An independent set of classifiers was trained for each speaker. In all cases, GMMs for particular prosodemes contained 5 mixtures. The classification results are presented in Table 2. Some speakers (namely M_{AJ}, M_{JS} and F_{KI}) have obviously and extraordinarily consistent speaking style because only a few individual prosodeme were classified as of a different type. The other speakers (M_{TJ} and F_{MR}) apparently often separated compound sentences into independent declarative phrases. This has 2 consequences:

1. Prosodemes P3-1 were classified as P1-1 because they actually correspond to that prosodeme type.
2. Prosodemes P1-1 were classified as P3-1 because training data for the P3-1 classifier contained a lot of P1-1 samples; therefore, the classifier was poorly trained.

Table 2. Classification of prosodemes in particular corpora (total numbers and percentages)

speaker	default prosodeme	classification			
		P1-1		P3-1	
M_{AJ}	P1-1	9,982	99.81 %	19	0.19 %
	P3-1	57	0.44 %	12,994	99.56 %
M_{JS}	P1-1	9,887	99.91 %	9	0.09 %
	P3-1	8	0.07 %	10,537	99.93 %
M_{TJ}	P1-1	9,075	91.70 %	821	8.30 %
	P3-1	3,516	20.12 %	12,994	79.88 %
F_{KI}	P1-1	9,884	99.87 %	13	0.13 %
	P3-1	25	0.76 %	3,537	99.24 %
F_{MR}	P1-1	9,523	96.41 %	355	3.59 %
	P3-1	1,198	13.96 %	7,383	86.04 %

Besides the GMM-based classification described in Section 3, some comparative experiments with support vector machines with various kernels [13] were also performed. The results were very similar and are therefore not presented in this paper. A more detailed classifier comparison is planned to be performed in our future work.

4.3 Listening Tests

The functionality of the proposed GMM-based classifiers was evaluated by listening tests. 10 participants took part in this listening test, most of them were speech processing experts who understood the theoretical background of the problem and had some former experience with listening tests.

The test contained 20 utterances for each speaker:

a) 5 phrases terminated with prosodeme P1-1,
b) 5 phrases terminated with prosodeme P3-1,
c) 10 phrases where the default prosodeme P3-1 was classified as P1-1; hereinafter, this prosodeme is denoted PX-Y.

The phrases a) and b) were selected to be prosodically unambiguous for the required prosodeme. Phrases c) were selected randomly. However, all the utterances were semantically neutral, i.e. the type of the phrase could not be determined from the text content.

Utterances of speakers M_{JS} and F_{KI} were not included in the test because of lack of phrases of type c). Thus, the test contained only utterances of 2 male speakers and one female speaker (M_{AJ}, M_{TJ} and F_{MR}, respectively).

The test results presented in Table 3 show that all the listeners were able to distinguish the prosodemes P1-1 and P3-1. Moreover, in most cases, they identified the prosodeme PX-Y as P1-1. That is in agreement with the results of the classifiers.

Table 3. Results of listening tests

speaker	prosodeme	listeners' decision [%]		
		P1-1	P3-1	undecided
M_{AJ}	P1-1	**100.0**	0.0	0.0
	P3-1	0.0	**94.0**	6.0
	PX-Y	**96.0**	0.0	4.0
M_{TJ}	P1-1	**98.0**	0.0	2.0
	P3-1	0.0	**100.0**	0.0
	PX-Y	**90.0**	1.0	9.0
F_{MR}	P1-1	**100.0**	0.0	0.0
	P3-1	2.0	**96.0**	2.0
	PX-Y	**100.0**	0.0	0.0
all	P1-1	**99.3**	0.0	0.7
	P3-1	0.7	**96.7**	2.7
	PX-Y	**95.3**	0.3	4.3

5 Conclusion

This paper presented some initial experiments on the detection of errors in the prosodic annotation of large speech corpora. The annotation errors are identified by simple GMM-based classifiers. Experiments performed on 5 large speech corpora revealed various numbers of suspicious prosodemes whose default prosodeme label did not match its new classification.

Listening test confirmed that the decision of the classifiers was correct in most cases and the new prosodeme label was correct. In the remaining cases, closer examination revealed two secondary causes of different classification: problems with the pitch extraction and problems with the default phonetic segmentation.

5.1 Future Work

In our future work, the experiments on classification of other types of prosodemes will be performed. By including the null prosodeme model, possibly incorrect phrase boundaries (shifted segmentation) could be also detected.

Given the promising initial results, the classifiers are planned to be used for an automatic correction of particular corpora. We expect that using corpora with corrected prosodeme labels for training a new voice in a TTS system should improve the overall quality of synthesized speech, especially its prosodic features.

Another aim is to develop speaker-independent classifiers that could be used for speech data from non-professional speakers whose speech prosody is not consistent enough to train new independent classifiers or the amount of speech data is low.

Last but not least, we intend to perform a more thorough comparison with other types of classifiers, e.g. support vector machines [13].

References

1. Hunt, A., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of ICASSP 1996, Atlanta, Georgia, pp. 373–376 (1996)
2. Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Communication 51, 1039–1064 (2009)
3. Ross, K., Ostendorf, M.: Prediction of abstract prosodic labels for speech synthesis. Computer Speech and Language 10, 155–185 (1996)
4. Toledano, D., Gómez, L., Grande, L.: Automatic Phonetic Segmentation. IEEE Transactions on Speech and Audio Processing 11(6), 617–625 (2003)
5. Wightman, C., Ostendorf, M.: Automatic labeling of prosodic patterns. IEEE Transactions on Speech and Audio Processing 2(4), 469–481 (1994)
6. Romportl, J., Matoušek, J., Tihelka, D.: Advanced Prosody Modelling. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 441–447. Springer, Heidelberg (2004)
7. Tihelka, D., Matoušek, J.: Unit Selection and its Relation to Symbolic Prosody: A New Approach. In: Proceedings of Interspeech 2006, Pittsburgh, Pennsylvania, USA, pp. 2042–2045 (2006)

8. Reynolds, D., Rose, R.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions Speech Audio Processing 3(1), 72–83 (1995)
9. Přibil, J., Přibilová, A.: Evaluation of influence of spectral and prosodic features on GMM classification of Czech and Slovak emotional speech. EURASIP Journal on Audio, Speech, and Music Processing 8, 1–22 (2013)
10. Talkin, D.: A Robust Algorithm for Pitch Tracking (RAPT). In: Kleijn, W.B., Paliwal, K.K. (eds.) Speech Coding and Synthesis, ch. 14, pp. 495–518. Elsevier Science (1995)
11. Speech Signal Processing Toolkit (SPTK), `http://sp-tk.sourceforge.net`
12. Matoušek, J., Tihelka, D., Romportl, J.: Building of a Speech Corpus Optimised for Unit Selection TTS Synthesis. In: Proc. of LREC 2008, Marrakech, Morocco (2008)
13. Vapnik, V.: Statistical Learning Theory. Wiley, Chichester (1998)