Detecting Emphasised Spoken Words by Considering Them Prosodic Outliers and Taking Advantage of HMM-Based TTS Framework

梁晖 *(Hui LIANG)*

Speech Processing Group, Computer Engineering & Networks Lab, ETH Zürich, Switzerland
liangh@tik.ee.ethz.ch

Abstract

A fresh approach to detecting emphasised spoken words, where the concept of one-class classification is adopted, is investigated in this research work, such that a major difficulty – collecting a large amount of well-annotated training data containing emphasis – can be avoided. The key idea, in brief, is that after rich context-dependent phone models are trained on common, neutrally read speech data in the HMM-based speech synthesis framework, emphasised words are considered prosodic outliers with respect to these "neutral" phone models and thus get detected. Experiments were conducted on speech data in the German language without any simplifying assumption (e.g. there was only one emphasised word in each utterance). Under many conditions this universally applicable approach was found to outperform totally random guessing, even though the emphasised words constituted only a small portion (i.e. 6.28%) of the test set.

Index Terms: emphasis detection, prosodic outlier, rich context modelling, HMM-based speech synthesis

1. Introduction

It is no longer a major challenge to generate sufficiently natural-sounding speech in the neutral reading style by the state-of-the-art statistical parametric techniques of text-to-speech synthesis [1, 2, 3]. Recently, speech scientists have been much interested in generating expressive prosody based on this technical framework of speech synthesis, for example, synthesising emotions (happiness, sadness, anger, etc) [4, 5] as well as emphasis, the latter being able to highlight the focus of an utterance and being a common phenomenon in spontaneous speech.

The intention and mood of a speaker determine which words in an utterance are emphasised. For example, any words in the sentence "*I didn't take the test yesterday*" may get emphasised in order to convey its speaker's particular subtext. Thus it is impossible for a speech synthesiser to predict when to synthesise an emphasis merely based on input plain text, unless some words are manually marked down as "to be emphasised" [6]. Speech-to-speech translation appears to be an important scenario where words to be emphasised are clearly marked down: words to be emphasised in output synthesised speech should be the translation of emphasised words in input natural speech. As a result, detecting emphasis in natural speech is an indispensable preliminary stage there.

To be clearer, *emphasis* in this research work only refers to audible prominence that carries a speaker's particular subtext [7]. It is distinct from other sorts of audible prominence such as lexical stress and pitch accent. There was a great deal of research conducted on detecting various kinds of audible prominence [8, 9, 10, 11, 12, 13, 14, 15, 16], ranging from simple ideas like counting the number of high-pitched frames [8] to statistical modelling like conditional random fields [9] and recurrent neural networks [10]. A common point amongst all these methods is that audible prominence was detected due to the prosodic contrast between prominence and the remainder within the same utterance. However, this contrast does not well suit emphasis detection: just to give a simple counterexample, words such as *all*, *never*, *any* generally sound prominent compared with neighbouring words, even if they do not carry any special subtext. Their sounding prominent is simply normal prosody. Hence emphasis detection should be based on a different prosodic contrast – the contrast between neutral and emphatic pronunciations of the same words. It is this contrast that reflects the nature of emphasis and that is effectively why humans can perceive emphasis.

Unfortunately there appears to be no big, well-annotated corpus containing emphasis, so it is difficult to model the contrast between emphasised words and their neutral pronunciations for emphasis detection. Recent inspirational research presented in [17] demonstrates that neutral and emotional speech could be classified according to their different projections onto reference bases trained only on neutral speech. This achievement based on the concept of one-class classification [18] suggests that emphasis may be detected given models trained only on neutrally read speech. There exist large speech corpora designed for speech recognition, in which many speakers were supposed to read prompts aloud in a *neutral* way. Therefore, it is expected in this research work that this kind of large corpus can result in models that well describe neutral speech, so that emphasised words can be viewed as outliers in prosody with respect to these models and then get detected. To the best of the author's knowledge, there has not been previous research on emphasis detection where the concept of one-class classification was adopted.

2. Proposed Method

The basic idea of the proposed method for this research work is to train a set of models on prosodic features extracted from a common, large, multi-speaker speech corpus designed for speech recognition and then to regard speech segments whose prosodic feature vectors do not fit corresponding models as components of an emphasised word. No big, well-annotated corpus containing emphasis is required as training data. This proposed method makes no simplifying assumption (for example, there is only one emphasised word in each utterance). Any spoken word in an utterance is considered possible to be emphasised. Therefore, the proposed method is universally applicable.

Since natural prosody varies according to different contexts, it makes more sense to model neutral prosodic features context by context. Rich context-dependent models are adopted in

HMM-based speech synthesis [1, 2] in order to reproduce extremely specific segmental and prosodic variations. Apart from that, the output speech of an HMM-based synthesiser sounds prosodically neutral when the training data is a large corpus designed for speech recognition [19]. Given these two facts, the training phase of the HMM-based speech synthesis framework is taken advantage of in the proposed method for the purpose of capturing the characteristics of context-dependent neutral prosodic patterns of natural speech.

2.1. Prosodic Features and Normalisation

After training data is segmented by forced alignment, phone duration, fundamental frequency (F0) and intensity are taken into account for preparing feature vectors. The technique for F0 extraction proposed in [20] is employed, as it provides continuous F0 contours. Original phone duration d_{phone}, F0 f_0 and intensity E are normalised as follows before being modelled:

$$\hat{d}_{\text{phone}} = \frac{d_{\text{phone}}}{\text{speaker-wise duration mean of all phones}},$$

$$\hat{f}_0 = \frac{f_0 - \text{sentence-wise F0 mean}}{\text{speaker-wise F0 standard deviation}},$$

$$\hat{E} = \frac{E - \text{sentence-wise intensity mean}}{\text{sentence-wise intensity standard deviation}}.$$

Then according to the phone-level forced alignment results, *one feature vector* is constructed for *each phone* with:

- $\hat{d}_{\text{phone}}$,
- $\hat{E}^{\text{max}}$: maximum of $\hat{E}$ in the period of length d_{phone},
- $\hat{f}_0^{\text{mean}}$, $\hat{f}_0^{\text{max}}$ and $\hat{f}_0^{\text{min}}$: mean, maximum and minimum of $\hat{f}_0$ in the period of length d_{phone}.

In addition, features derived from continuous wavelet transformation (CWT)-based decomposition of *original* continuous F0 contours are also taken into consideration. CWT has been recently proposed to model F0 in the context of speech synthesis [21]. It was shown that the systems using CWT-based F0 decomposition tended to outperform those where F0 was modelled directly. According to [21], the continuous wavelet transform of an F0 contour f_0 is defined as

$$W^{(f_0)}(\tau, t) = \tau^{-1/2} \int_{-\infty}^{\infty} f_0(x)\psi\left(\frac{x-t}{\tau}\right) \mathrm{d}x \quad (\tau > 0),$$

where $\psi(\cdot)$ denotes a mother wavelet of the Mexican hat, and the contour f_0 can be reconstructed as

$$f_0(x) = \int_{-\infty}^{\infty} \int_{0}^{\infty} W^{(f_0)}(\tau, t)\tau^{-5/2}\psi\left(\frac{x-t}{\tau}\right) \mathrm{d}\tau \, \mathrm{d}t.$$

The decomposition and reconstruction may be approximated by choosing 10 scales of differing frequency, one octave apart. The 10 scales as ten separate streams are calculated as follows:

$$W_i^{(f_0)}(t) = W^{(f_0)}(2^{i+1}\tau_0, t)(i + 2.5)^{-5/2}, \tag{1}$$

where i = 1, 2, ..., 10 and τ_0 = 5ms. The contour f_0 is approximately recovered through the ad-hoc formula

$$f_0(t) = \sum_{i=1}^{10} W_i^{(f_0)}(t) + \epsilon(t),$$

where $\epsilon(t)$ is reconstruction error. Eventually as per the phone-level forced alignment results, the feature vector for each phone also includes:

- w_4^{mean}, w_4^{max}, w_4^{min} (i = 7 in Eq. (1)) and w_5^{mean}, w_5^{max}, w_5^{min} (i = 6 in Eq. (1)): mean, maximum and minimum of the 4th and 5th scales of the CWT-based decomposition result in the period of length d_{phone}.

Ribeiro et al [22] discovered that the distributions of intonational phrases and content words matched those of local maxima of scales 4 and 5, respectively. This is the reason why these two scales are employed as components of the feature vectors in the proposed method.

In summary, there are 11 dimensions in total in every prosodic feature vector. Because of the normalisation, prosodic feature vectors from different training speakers are treated as if they were from the same person.

2.2. Context-Dependent Modelling of Neutral Prosody

All the contextual factors used in a typical HMM-based speech synthesiser (see [1] and [2] for details), which involve the five levels ranging from phones to utterances, are employed in the proposed method. Each rich context-dependent phone model derived from the above-mentioned, 11-dimensional prosodic feature vectors is in effect formed by a single multivariate Gaussian distribution with a diagonal covariance matrix. Owing to the lack of sufficient training data for the huge amount of combinations of the contextual factors' values, decision tree-based clustering [23] is applied so as to tie these rich context-dependent Gaussian distributions.

2.3. Finding Emphasised Words in Test Data

The likelihood of each feature vector of training data given tied rich context-dependent phone models is calculated first of all, in order that likelihood thresholds for detecting outliers (i.e. components of emphasised words) can be determined. Then the likelihoods of feature vectors of test data (extracted as described in section 2.1) given these tied rich context-dependent phone models are calculated. Whether phones in the test data are components of emphasised words can be decided according to the likelihood thresholds.

Finally, whether a word is emphasised can be decided based on the decisions at the phone level according to certain criteria: for example, a word is emphasised when the vowel carrying the primary lexical stress in the word is considered emphasised, or when there are more phones considered emphasised than phones considered neutral in the word.

3. Experiments

3.1. Speech data

All the experiments were conducted in the German language. The training data was corpora PHONDAT1 and PHONDAT2 [24], amounting to 18972 utterances and 217 speakers. The test data included 130 utterances and 4 speakers, which were selected from a pilot multilingual corpus recorded at the University of Geneva [25]. Details may be found in Table 1.

Table 1: Details of the test set

Speaker	# of words	# of emph. words	Percentage
C1_12	414	30	7.25%
B1_06	260	15	5.77%
B2_11	250	14	5.60%
B1_13	238	14	5.88%
Total	1162	73	6.28%

Please note that these low percentages are not surprising, since typically emphasis appears just a couple of times in one utterance and does not appear in every utterance.

3.2. System Description

Since the number of tied rich context-dependent phone models can be controlled in the MDL criterion [26] in decision tree-based clustering, 15 different speaker-independent systems were built upon the same training data such that their tied phone model sets reached various extent of generalisation. Only all the 7104 rich context-dependent vowels were modelled in the 15 systems. The consonants did not contribute to the detection of any emphasised word. Values of the MDL factor [27] chosen during training and their resulting numbers of tied phone models are listed in Table 2.

Table 2: Details of the 15 speaker-independent systems

System	MDL factor	# of tied phone models	Compression ratio after clustering (r_{clst})
I	0.01	7104	100%
II	0.02	7092	99.83%
III	0.05	6937	97.65%
IV	0.1	6368	89.64%
V	0.2	4995	70.31%
VI	0.3	3943	55.50%
VII	0.4	3215	45.26%
VIII	0.5	2680	37.73%
IX	0.6	2281	32.11%
X	0.8	1758	24.75%
XI	1.0	1357	19.10%
XII	1.5	827	11.64%
XIII	2.0	560	7.88%
XIV	3.0	297	4.18%
XV	5.0	131	1.84%

3.3. Measures of Performance

The following measures of performance of detection of emphasised *words* were employed for assessing the 15 systems:

	predicted neu.	predicted emph.
true neu.	a	α
true emph.	β	b

- Recall (rec.) $= b \ / \ (b + \beta)$
- Precision (pre.) $= b \ / \ (b + \alpha)$
- Accuracy (acc.) $= (a + b) \ / \ (a + \alpha + b + \beta)$

In the case of totally random guessing, both recall and accuracy are 50%, and precision is the percentage of emphasised words in test data (i.e. 6.28% in this paper as per Table 1).

3.4. System Performance and Observations

German is a language where every word contains lexical stress. Generally speaking, the vowel carrying the primary lexical stress of a word is emphasised when one wants to emphasise the word, although occasionally one emphasises the unstressed negative prefixes of some words on purpose. Hence in the following experiments, *whether a word was emphasised* was determined by the decision as to whether the vowel carrying the primary lexical stress in the word was considered emphasised.

3.4.1. General performance of detection

The likelihood threshold for every tied phone model to detect prosodic outliers amongst all its associated feature vectors can be automatically chosen, such that certain accuracy of detection upon the entire training data (hereafter k_{trn}) is achieved. In the experiments in this research work, k_{trn} equalled 100%, 95%, 90%, 85%, 80%, 75%, 70% and 65% in turn. These values of k_{trn} reflected various extent of supplementary generalisation provided to the 15 tied phone model sets. Fig. 1 shows the performance of the 15 systems in terms of recall, precision and accuracy at the word level under different conditions.

First of all, it can be observed in Fig. 1 that k_{trn} needed to remain in a narrow range roughly from 90% to 95% in order that most of the 15 systems outperformed totally random guessing. In other words, this range of k_{trn} could help to lead to the best generalisation of the tied model sets. Furthermore, for the purpose of achieving relatively good performance, Fig. 1 indicates that the number of the tied phone models should be reduced to around 20% to 30% of that of the original rich context-dependent phone models (this region of r_{clst} is denoted by vertical dashed black lines in Fig. 1). In other words, the MDL factor ought to be set to around 0.6 to 1.0. These two ranges, unsurprisingly, may not be used directly in a detector to be trained on other speech data and to be applied to other test set, but would be capable of serving as a sensible frame of reference for building such a detector.

Compared with the recall in Fig. 1(a) and the accuracy in Fig. 1(c), the precision in Fig. 1(b) was indeed low. However, the precision always remained above the chance level except a handful of extreme cases. This is a signal that the proposed method rests upon an idea on the right track. Apart from that, the relative increment of precision from the chance level could reach 110.0% to 155.3% (see systems IX, X and XI with k_{trn} =

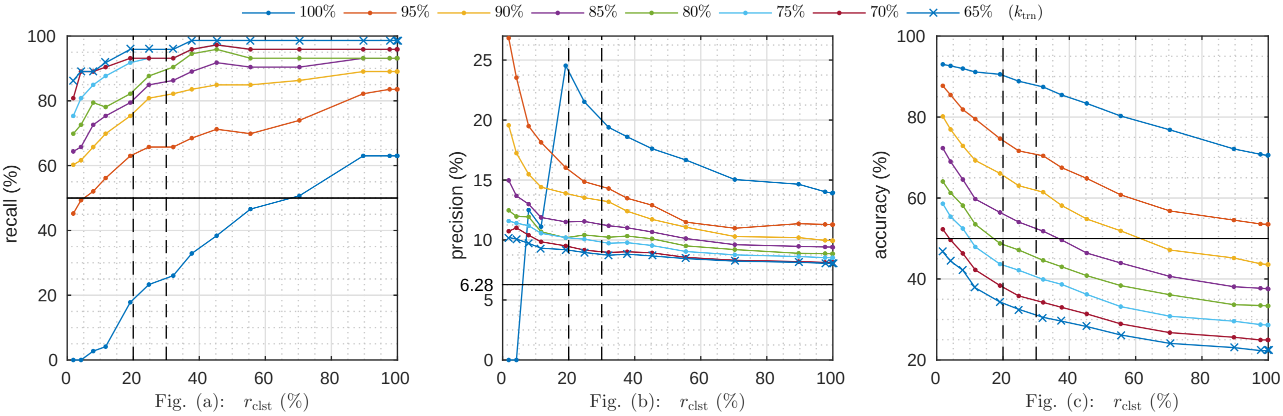

Fig. (a): r_{clst} (%)

Fig. (b): r_{clst} (%)

Fig. (c): r_{clst} (%)

Figure 1: *Word-level* recall, precision and accuracy of the 15 systems obtained upon the test data, the horizontal solid black lines denoting the case of totally random guessing

90% and 95%), which were effectively substantial amounts.

3.4.2. Impact of individual components of feature vectors

As per the findings in section 3.4.1, only three (IX, X, XI) out of the 15 systems with k_{trn} = 90% and 95% were taken a closer look at so as to investigate the individual impact of each component of the feature vectors. To be more specific, the tied phone models of systems IX, X and XI as well as the 11-dimensional feature vectors of the training and test data were used dimension by dimension to calculate all the likelihoods required for detecting outliers in the proposed method. Results showing this component-specific impact are presented in Tables 3, 4 and 5. The maximum in each column is marked in bold type.

Table 3: Performance of system IX (in percentage)

	k_{trn} = 90%			k_{trn} = 95%		
	rec.	pre.	acc.	rec.	pre.	acc.
$\hat{d}_{\text{phone}}$	47.95	12.11	72.40	42.47	13.72	77.60
$\hat{E}^{\text{max}}$	**82.19**	11.63	55.67	**79.45**	12.86	61.44
$\hat{f}_0^{\text{mean}}$	54.79	13.29	72.21	50.68	16.89	79.40
$\hat{f}_0^{\text{max}}$	61.64	14.80	72.87	50.68	17.29	79.87
$\hat{f}_0^{\text{min}}$	52.05	12.84	72.31	46.58	14.78	77.79
w_4^{mean}	42.47	12.02	74.57	28.77	12.28	80.91
w_4^{max}	36.99	10.67	74.29	28.77	12.21	80.81
w_4^{min}	36.99	10.80	74.57	23.29	10.30	80.72
w_5^{mean}	47.95	14.29	76.56	39.73	**18.47**	83.74
w_5^{max}	52.05	**15.64**	77.32	35.62	16.05	82.70
w_5^{min}	42.47	14.16	**78.26**	32.88	17.39	**84.59**
all	82.19	13.19	61.44	65.75	14.29	70.42

Table 4: Performance of system X (in percentage)

	k_{trn} = 90%			k_{trn} = 95%		
	rec.	pre.	acc.	rec.	pre.	acc.
$\hat{d}_{\text{phone}}$	45.21	12.22	73.79	41.10	14.63	79.38
$\hat{E}^{\text{max}}$	**82.19**	11.76	56.24	**76.71**	12.56	61.53
$\hat{f}_0^{\text{mean}}$	53.42	13.73	73.63	49.32	17.31	80.25
$\hat{f}_0^{\text{max}}$	60.27	15.28	74.20	49.32	18.37	81.38
$\hat{f}_0^{\text{min}}$	50.68	13.17	73.53	42.47	15.05	79.49
w_4^{mean}	41.10	12.10	75.33	28.77	12.50	81.19
w_4^{max}	36.99	11.02	75.05	28.77	12.80	81.57
w_4^{min}	35.62	10.92	75.52	23.29	10.30	80.72
w_5^{mean}	46.58	14.53	77.41	39.73	**18.59**	83.84
w_5^{max}	49.32	**15.38**	77.79	36.99	16.56	82.80
w_5^{min}	41.10	14.22	**78.83**	30.14	16.30	**84.50**
all	80.82	13.53	63.04	65.75	14.86	71.64

Table 5: Performance of system XI (in percentage)

	k_{trn} = 90%			k_{trn} = 95%		
	rec.	pre.	acc.	rec.	pre.	acc.
$\hat{d}_{\text{phone}}$	43.84	12.96	75.78	41.10	16.39	81.46
$\hat{E}^{\text{max}}$	**80.82**	12.02	57.84	**75.34**	13.10	63.80
$\hat{f}_0^{\text{mean}}$	52.05	14.02	74.67	46.58	18.78	82.42
$\hat{f}_0^{\text{max}}$	57.53	**15.44**	75.33	50.68	**20.67**	83.18
$\hat{f}_0^{\text{min}}$	47.95	13.21	74.67	35.62	14.21	80.72
w_4^{mean}	39.73	12.08	75.90	24.66	11.46	81.66
w_4^{max}	32.88	10.53	76.09	24.66	12.08	82.42
w_4^{min}	31.51	10.41	76.56	19.18	9.09	81.19
w_5^{mean}	41.10	14.42	79.11	34.25	18.66	85.16
w_5^{max}	42.47	14.83	79.21	32.88	16.67	84.03
w_5^{min}	32.88	12.90	**80.06**	26.03	16.81	**86.01**
all	75.34	13.89	66.07	63.01	16.03	74.67

The most noticeable thing in these three tables would be the high recall that $\hat{E}^{\text{max}}$ led to in all the six cases. It is therefore clear that $\hat{E}^{\text{max}}$ is a distinctive feature of emphasised words in German. This makes sense since the stressed syllable of an emphasised word typically sounds louder, if lexical stress exists in the language. However, $\hat{E}^{\text{max}}$ is merely discriminative to an extent – the precision was 5.35 to 6.82 percentage points higher than the chance level 6.28% though it was roughly doubled.

The three tables show that w_4^{mean}, w_4^{max} and w_4^{min} were the least useful amongst the pitch-related feature vector components. $\hat{f}_0^{\text{mean}}$, $\hat{f}_0^{\text{max}}$ and $\hat{f}_0^{\text{min}}$ led to much higher recall than w_5^{mean}, w_5^{max} and w_5^{min}, while w_5^{mean}, w_5^{max} and w_5^{min} produced comparable precision and slightly higher accuracy. Even though scales 4 and 5 of the CWT-based F0 decomposition result had certain physical meanings [22], they didn't appear to be helpful individually as expected. $\hat{f}_0^{\text{max}}$ was arguably the best amongst the pitch-related feature vector components in this proposed method.

$\hat{d}_{\text{phone}}$ actually obeyed the gamma distribution rather than the Gaussian distribution. The Gaussian distribution is employed in HMM-based speech synthesis to model every kind of feature (including duration), so modelling $\hat{d}_{\text{phone}}$ by the Gaussian distribution was tried in this research work. According to Tables 3, 4 and 5, no evidence indicates that this approximation was unfavourable.

4. Discussion

The previous analysis focused on the combinations of r_{clst} and k_{trn} that resulted in all the three measures of performance (recall, precision and accuracy) being above the chance level. During the practical use of a speech-to-speech translator that can transfer emphasis across languages, not conveying a (correct) subtext could be preferable to conveying a wrong subtext. In other words, recall may be sacrificed for precision (and accuracy). According to Fig. 1, r_{clst} should be very small and k_{trn} should fall into the range of 95% to 100% in order to fulfil this practical requirement. In this case, the precision may be expected to be around 4.3 times as high as the chance level 6.28%.

Each sentence in PHONDAT1 and PHONDAT2 was read by quite a few speakers. This is advantageous to training reliable rich context-dependent phone models, but the number of different contexts is relatively limited (compared with English corpora like WSJ0 and WSJCAM0), which is disadvantageous to capturing the characteristics of context-dependent neutral prosodic patterns of natural speech. However, given the aforementioned performance, it would be reasonable to anticipate better performance produced by the proposed method when training data with much more prosodic contexts is available.

5. Conclusions

The detection of emphasised words is investigated in this research work from a fresh angle in order to avoid the difficulty in collecting a large, well-annotated corpus containing emphasis. The proposed method is based on the concept of one-class classification and the intrinsic prosodic contrast between neutral and emphatic pronunciations of the same words. Duration, F0, intensity and CWT-based decomposition of F0 contours contribute to the construction of feature vectors. Rich context-dependent modelling and decision tree-based clustering are employed to model the "neutral feature space" such that emphasised words can be found as prosodic outliers. The experimental results show that the proposed method outperformed totally random guessing under many conditions, even though the emphasised words constituted only 6.28% of the test set. This is a clear sign that the proposed method rests upon an idea on the right track. Since no simplifying assumption is involved, the proposed method is applicable to any scenario of emphasised word detection.

6. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[2] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. of the IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[3] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, May 2015.

[4] J. Latorre, V. Wan, M. J. F. Gales, L. Chen, K. K. Chin, K. Knill, and M. Akamine, "Speech factorization for HMM-TTS based on cluster adaptive training," in *Proc. of Interspeech*, Sep. 2012, pp. 971–974.

[5] L. Chen and N. Braunschweiler, "Unsupervised speaker and expression factorization for multi-speaker expressive synthesis of ebooks," in *Proc. of Interspeech*, Aug. 2013, pp. 1042–1046.

[6] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in HMM-based speech synthesis," in *Proc. of ICASSP*, Mar. 2010, pp. 4238–4241.

[7] K. J. Kohler, "What is emphasis and how is it coded?" in *Proc. of Speech Prosody*, May 2006, pp. 748–751.

[8] L. S. Kennedy and D. P. W. Ellis, "Pitch-based emphasis detection for characterization of meeting recordings," in *Proc. of ASRU*, Dec. 2003, pp. 243–248.

[9] V. K. Rangarajan Sridhar, A. Nenkova, S. Narayanan, and D. Jurafsky, "Detecting prominence in conversational speech: pitch accent, givenness and focus," in *Proc. of Speech Prosody*, May 2008, pp. 453–456.

[10] Y. Ren, S.-S. Kim, M. Hasegawa-Johnson, and J. Cole, "Speaker-independent automatic detection of pitch accent," in *Proc. of Speech Prosody*, Mar. 2004, pp. 521–524.

[11] M. Heldner, E. Strangert, and T. Deschamps, "A focus detector using overall intensity and high frequency emphasis," in *Proc. of ICPhS*, 1999, pp. 1491–1494.

[12] P. Martin, "Prominence detection without syllabic segmentation," in *Proc. of Speech Prosody*, May 2010.

[13] B. Arons, "Pitch-based emphasis detection for segmenting speech recordings," in *Proc. of ICSLP*, vol. 4, Sep. 1994, pp. 1931–1934.

[14] F. Tamburini and C. Caini, "An automatic system for detecting prosodic prominence in American English continuous speech," *International Journal of Speech Technology*, vol. 8, no. 1, pp. 33–44, 2005.

[15] M. Avanzi, A. Lacheret-Dujour, and B. Victorri, "A corpus-based learning method for prominence detection in spontaneous speech," in *Proc. of Speech Prosody*, May 2010.

[16] J. M. Brenier, D. M. Cer, and D. Jurafsky, "The detection of emphatic words using acoustic and lexical features," in *Proc. of Eurospeech*, Sep. 2005, pp. 3297–3300.

[17] J. P. Arias, C. Busso, and N. Becerra Yoma, "Energy and F0 contour modeling with functional data analysis for emotional speech detection," in *Proc. of Interspeech*, Aug. 2013, pp. 2871–2875.

[18] D. M. J. Tax, "One-class classification: Concept-learning in the absence of counter-examples," Ph.D. dissertation, Delft University of Technology, Jun. 2001.

[19] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, Y. Guan, R. Hu, K. Oura, Y.-J. Wu, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis – Analysis and application of TTS systems built on various ASR corpora," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 984–1004, Jul. 2010.

[20] T. Ewender, S. Hoffmann, and B. Pfister, "Nearly perfect detection of continuous F0 contour and frame classification for TTS synthesis," in *Proc. of Interspeech*, Sep. 2009, pp. 100–103.

[21] A. Suni, D. Aalto, T. Raitio, P. Alku, and M. Vainio, "Wavelets for intonation modeling in HMM speech synthesis," in *Proc. the 8th ISCA Speech Synthesis Workshop*, Aug. 2013, pp. 285–290.

[22] M. S. Ribeiro, J. Yamagishi, and R. A. J. Clark, "A perceptual investigation of wavelet-based decomposition of F0 for text-to-speech synthesis," in *Proc. of Interspeech*, Sep. 2015, pp. 1586–1590.

[23] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. of the Workshop on Human Language Technology*, 1994, pp. 307–312.

[24] C. Draxler, "Introduction to the Verbmobil-PhonDat database of spoken German," in *Proc. of the 3rd International Conference on the Practical Application of PROLOG*, Apr. 1995, pp. 201–212.

[25] P. N. Garner, R. Clark, J.-P. Goldman, P.-E. Honnet, M. Ivanova, A. Lazaridis, H. Liang, B. Pfister, M. S. Ribeiro, E. Wehrli, and J. Yamagishi, "Translation and prosody in Swiss languages," in *Nouveaux cahiers de linguistique française 31*, Sep. 2014, pp. 211–221.

[26] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *Journal of Acoustical Society of Japan (E)*, vol. 21, no. 2, pp. 79–86, 2000.

[27] K. Oura, "List of modifications made in HTS (for version 2.2)," Jul. 2011. [Online]. Available: http://hts.sp.nitech.ac.jp/archives/2.2/HTS_Document.pdf

****

The research work presented in this paper was funded by the SIWIS project (SNSF Grant CRSII2-141903).

The author would like to thank Manuel Sam Ribeiro at the Centre for Speech Technology Research, University of Edinburgh for providing the F0 decomposition result of the training and test sets.