Emphasis Recreation for TTS using Intonation Atoms

Pierre-Edouard Honnet[1,2] *and Philip N. Garner*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne, Switzerland
{pierre-edouard.honnet, phil.garner}@idiap.ch

Abstract

We are interested in emphasis for text to speech synthesis. In speech to speech translation, emphasising the correct words is important to convey the underlying meaning of a message. In this paper, we propose to use a generalised command-response (CR) model of intonation to generate emphasis in synthetic speech. We first analyse the differences in the model parameters between emphasised words in an acted emphasis scenario and their neutral counterpart. We investigate word level intonation modelling using simple random forest as a basis framework, to predict the parameters of the model in the specific case of emphasised word. Based on the linguistic context of the words we want to emphasise, we attempt at recovering emphasis pattern in the intonation in originally neutral synthetic speech by generating word-level model parameters with similar context. The method is presented and initial results are given, on synthetic speech.

Index Terms: Intonation, emphasis, generalised command-response model, random forest, text-to-speech synthesis

1. Introduction

Although it has been investigated for decades, the interest for speech-to-speech translation (S2ST) is still growing. The state of the art systems are built around three main components: automatic speech recognition (ASR), automatic machine translation (MT), and text to speech (TTS) synthesis, with the output of each subsystem simply pipelined into the next one. To improve human-human interaction in the cross lingual context, the system should be able to transfer the non verbal intentions of participants, which implies translating and synthesisng more than just the recognised text.

In a spoken sentence, the speaker tends to emphasise some words, in order to draw the attention of the listener to these words. Emphasising different words can also change the underlying meaning of the sentence. Tsiartas *et al.* [1] conducted a large-scale human evaluation on the perception of S2ST quality and showed that the perceived quality of S2ST was correlated with cross-lingual prosodic emphatic transfer. In other words, emphasising the correct words in the output language in TTS based on the emphasised words in the input language helps in the S2ST task.

Although there has been some work on the personalisation of TTS for S2ST systems in the last decade, with some projects such as EMIME[1] [2], there is still relatively little work on the improvement of TTS systems in the context of S2ST. Parlikar *et al.* [3] worked on improving TTS where the input of the system is the output of the translation module. They proposed to insert pauses, to replace untranslated words with fillers and to use alternate translation to minimise the cost of their unit selection system to make the speech more intelligible. Another aspect of S2ST that deserves some improvement is the transfer of speakers' intentions. Anumanchipalli *et al.* [4] recently proposed to translate the emphasis in S2ST. More recently, Do *et al.* [5] proposed to model word-level emphasis and use conditional random fields to translate emphasis to a target language. Pause prediction for improving emphasis in S2ST was also investigated by Do and colleagues [6].

More generally, emphasis has received some attention in the TTS context. For instance, Yu *et al.* [7] proposed to model word-level emphasis in the context of HMM-based emphasis, using different decision tree clustering techniques. Another approach, proposed by Hirose *et al.* [8], consisted of post-processing the F_0 contour with the command-response model of Fujisaki [9].

In our recent work, we investigated the transfer of some local intonation components to mimic emphasis on a target word in a neutral sentence [10]. After analysing differences between neutral and emphasised scenarios, we identified the manifestation of emphasis to be correlated with a higher number of local components in the intonation contour for the same word in the same context. We showed that transferring the most prominent local components (both positive and negative) to the F_0 contour of a neutral word elicits the perception of emphasis in an originally neutral sentence.

In this paper, we investigate the use of clustering methods to predict the local intonation components of emphasised words. Using emphasised word F_0 decomposition in context, we attempt to predict the model parameters for an emphasised word in some specific context. These components can then be used as word-level intonation. This work is an initial investigation of intonation modelling for emphasised words. On the synthesis aspect, this could be used with some complementary method, like duration alteration, or intensity modification. In the more general framework of translating emphasis in S2ST, some emphasis detection system (e.g. the recent work of Cernak and colleagues [11, 12]) can be used to provide the machine translation additional information, which can further be transmitted to the TTS system. For this study, we restrict ourselves to the intra-lingual case, but due to the language independence of the intonation model used, it seems reasonable to assume that this method can work for any given language.

We first present the intonation model which is used for this work and confirm the intuition that it suits emphasis transfer with mutual information analyses. Later, we describe our framework to predict word-level intonation in the case of emphasised word in a sentence. Some initial results are finally presented with perspectives on how to exploit the model.

[1]http://www.emime.org/

2. Generalised Command-Response model

2.1. Related work

The literature provides a lot of work in intonation modelling. There are various categories of models, with different applications. The state of the art F_0 generation for speech synthesis simply follows the way other acoustic features are generated, using hidden Markov models (HMMs) [13], or more recently deep neural networks (DNNs) [14]. In these frameworks, the intonation is predicted frame by frame and relies on the linguistic context given in the input of the system.

Some of the best known external models are reviewed in our previous work [15, 16]. Fujisaki and colleagues have worked for several decades on a model which tries to model the underlying process of human intonation production [9, 17, 18]. One of its applications is style adaptation: by modifying the commands of the model in the F_0 produced by the TTS models, the authors control the prosody of the synthetic speech [8]. In a similar fashion, the CR model was used for intonation contour reshaping to add focus in the synthetic speech [19]. The CR model was also implemented as an intonation generation model using specific topology hidden Markov models [20, 21].

Anumanchipalli *et al.* [4] exploited the *tilt* model [22] to train a conversion function between vectors from input and output languages from a parallel corpus.

2.2. Generalised command-response approach

The CR model is attractive due to its physiological basis, which makes it theoretically language independent. While Fujisaki [18] relates two types of components to two muscle actions, Strik [23] advocates that more muscles play a role in the control of the vocal fold tension, and that the subglottal pressure is also responsible for F_0 variations.

We proposed the generalised command-response (GCR) model as an alternative command-response model characterised by an automatic parameter extraction procedure [15]. The decomposition of the contour is based on the matching pursuit algorithm with a dictionary of higher-order system impulse responses of the form of $G_{k,\theta}(t)$ (1), that happen to have the same functional form as a gamma distribution:

$$G_{k,\theta}(t) = \frac{1}{\theta^k \Gamma(k)} t^{k-1} e^{-t/\theta} \quad \text{for} \quad t \geq 0 \qquad (1)$$

where k is the order of the model (the shape), and θ the scale, Γ is the gamma function.

The model has two types of components, global (for long term variations) and local. We further improved the perceptual relevance of the elements that are extracted from the F_0 contour by using a weighted correlation as a cost function based on energy and probability of voicing ($w(i) = e(i) * p(i)$ where $e(i)$ and $p(i)$ are respectively the energy and probability of voicing of frame i). Using this perceptually relevant measure then allows to extract components which are not only strongly correlated with the F_0 based on raw magnitude.

Additionally, we use a different global component shape, similar to (1) with higher values for θ. For more details, see [15, 16].

The model parameters given by the decomposition are then for each local and global component – that we call atom – a position, amplitude and θ. The system order, k in (1), is fixed as we assume the same order for the different impulse responses. The complete contour is then reconstructed as:

$$F_0 = G_{\text{phrase},k,\theta}(t - t_{\text{phrase}}) + \sum_{n=1}^{N} A_n G_{k_n,\theta_n}(t - t_n) \qquad (2)$$

where t_{phrase} is the position of the phrase component, k and θ its order and scale, N the number of atoms, t_n is the position of the local component n, A_n its amplitude, and k_n and θ_n its order and scale.

2.3. Relevance of GCR features

To assess the relevance of the parameters extracted from our model, we examined the mutual information shared between the parameters and some linguistic features. By looking at mutual information, we expect to find some clues on how atom parameters relate with linguistic features. We measured the mutual information between atom parameters (amplitude, position and θ) and classical contextual features, and looked at the differences between neutral and emphatic data. Our hypothesis is that when a word is emphasised, the model will extract different types of atoms, then we should observe a higher mutual information between emphasis and atom parameters. If we denote the labels as L and the model features as F_i, the mutual information was calculated the following way:

$$I(L, F_i) = \sum_{l} \sum_{f \in F_i} p(l, f) \log_2 \left(\frac{p(l, f)}{p(l)p(f)} \right) \qquad (3)$$

with $p(l, f)$ the joint probability of L and F_i, and $p(l)$ and $p(f)$ their respective marginal probabilities. These probabilities are the maximum likelihood estimate based on occurrences in the data. The model parameters were quantised as follows: between 0 and 10 for position (relative position in the syllable) yielding 11 possible values, and between 0 and 9 for amplitude yielding 11 possible values. θ's range was 0.01 - 0.05 with 0.005 steps, yielding 9 possible values. The labels l are binary (accented or not, stressed or not, emphasised or not). The results presented are normalised by the entropy of the labels:

$$H(L) = -\sum_{l} p(l) \log_2(p(l)) \qquad (4)$$

Table 1 shows normalised mutual information in the case of a single English female speaker, for about 300 neutral read sentences.

The results indicate that mutual information between amplitude and accent and between relative position in the syllable and accent are the highest for single feature and single context. As can be expected, the syllable which are both accented and stressed have higher mutual information with atom parameters. We also notice that using all atom parameters (amplitude, position and θ) does not bring more information than using position and amplitude only.

Given these initial observations, we looked at the mutual information between amplitude, position and number of atoms per syllable, and accent, stress and emphasis. In that case, the data consisted of about 300 sentences from multiple English speakers, in two scenarios: neutral sentence and sentence with one emphasised word or group of words. To compare both cases, the same *"target words"* were used: the word emphasised in the emphasised case was selected as *target word*, and in the neutral case it was also tagged as emphasised, to see its effect on the parameters. The results are presented in table 2.

Table 1: Normalised mutual information between atoms and linguistic features, for *Nancy* .

Context/Feats	Amp	θ	Amp, θ	Pos	Pos, Amp	Pos, θ	Pos, Amp, θ
Accent	11.1	8.7	13	11.1	23.3	22.6	23.3
Stress	8.2	6.8	9.5	8.1	22.3	21.8	22.2
Acc. & Stress	13.4	11	16	13.9	27.1	26.6	27.3
Acc. or Stress	7.9	6.2	9.1	7.7	23.8	23.2	23.7

Table 2: Normalised mutual information between atoms and linguistic features [neutral / emphasised].

Context/Feats.	Amp.	Pos.	N_{atoms} in syllable
Accent	12.4 / 13.0	14.1 / 14.9	8.4 / 8.8
Stress	10.3 / 10.4	11.4 / 11.5	7.3 / 7.8
Emphasis	20.8 / 17.4	24.0 / 20.6	11.9 / **18.9**
Acc. & Stress	15.6 / 16.0	18.0 / 18.8	10.3 / 10.8
Emph. & Stress	40.5 / 29.8	48.3 / 35.2	26.6 / **44.4**
Emph. & Acc.	53.8 / 47.5	60.4 / 55.8	38.2 / **56.1**

Table 3: Normalised mutual information between atoms and linguistic features [neutral / emphasised] – same number of atoms.

Context/Feats.	Amp.	Pos.	θ
Accent	12.5 / 13	14.3 / 14.8	9.9 / 10.6
Stress	10.2 / 10.3	11.3 / 11.8	8.5 / 8.9
Emphasis	24.6 / **25.2**	28.5 / **29.8**	21.7 / **23.7**
Acc. & Stress	15.7 / 16.4	18.3 / 18.9	12.1 / 12.9
Emph. & Stress	48 / **51.9**	57.8 / **60**	41.7 / **46.9**
Emph. & Acc.	65.6 / **69.9**	74.9 / **78.1**	60 / **62.8**

In this table, we can see that the most discriminant feature to distinguish emphasised and neutral data is the number of atoms in the syllable. This contradicts our hypothesis that the atoms in an emphasised word are different than the ones in a neutral word. One possible explanation for this finding is that the F_0 curve presents more variations in the region of emphasised word, resulting in a need for more atoms to fit the curve. To verify further the difference between the "principal" atoms in each word, we looked at the same measures as in table 2, but with a constraint on the number of atoms: we selected only the first n atoms – ranked by amplitude – in the emphasised word, where n is the number of atoms in the same target word in the neutral case. In the cases where the neutral version had more atoms, its number was restricted in the same way, to always have the same number of atoms. Table 3 gives the results for mutual information with the same number of atoms.

We can see that when the number of atoms is the same for the neutral and emphasised case, mutual information between both amplitude and position and accent, stress and emphasis is higher in the emphasised case. This is particularly true for emphasis. Then, in addition to the fact that emphasis manifests itself with more atoms, emphasis seems to be expressed through different patterns for the components: different positions, amplitudes and θ. It validates our intuition, that when a word is emphasised, the components resulting from the decomposition are distinguishable from the neutral case.

However, it is not obvious how intonation is affected by the presence of emphasis on a particular word. This means that, if we use the GCR approach to model the F_0, simply modifying the amplitude or θ of the atoms, i.e transforming them into bigger atoms, is unlikely to produce emphasis. Previous experiments showed that when increasing the amplitude of the atoms, differences were noticeable, but the words for which atom amplitudes were altered were not perceived as emphasised. For this reason, the next section presents a way of predicting atoms in an emphasised scenario.

3. Modelling word level intonation in emphasised scenario

3.1. Application of GCR model to emphasis transfer

The GCR model lends itself to emphasis transfer. Using an emphasis detection module combined with ASR-based automatic time alignment, it is possible to identify which word is emphasised in a sentence and its boundaries (we do not tackle this problem in this work; it can be solved using different methods, e.g. [11, 12]).

In our previous work [10], given parallel data including neutral and emphasised speech, by retrieving the parameters of our model for both cases, we showed that adding the most prominent atoms from an emphatic word in a neutral sentence consistently increased the perception of emphasis on the target word. Subjective tests confirmed that listeners were able to identify artificially emphasised words in most cases.

3.2. Random forest based word level intonation

In this paper, based on the observation made on mutual information when using a constrained number of components to model word level intonation, we investigate the possibility of predicting model parameters in the context of emphasised words. While regression trees offer a simple but powerful way of clustering data based on linguistic features, random forest present the advantage of a better generalisation over the data. Using random subset of the data for training multiple trees allows to create independent trees, which then are all used to give the output values. Then, we hypothesise that random forests (RF) can predict the model parameters for word level emphasis. Given that the parameters of the model correlate more with emphasised speech than neutral speech, we investigate whether we can learn these parameters in an emphasised case. Initial experiments showed that random forest were better suited to this task, compared to decision trees.

To predict the word level parameters of the GCR model, we propose to use a subset of the contextual features generally used for HMM-based speech synthesis [13]:

- Number of syllables in the word
- gPOS (guess part of speech) of the word
- Stress position(s) in the word [0-3]
- Accent position(s) in the word [0-2]
- Word position in phrase
- TOBI endtone of the phrase

- Phrase position in utterance
- Number of phrases in utterance
- Number of words in utterance
- Number of syllables in utterance

Based on some preliminary observations to characterise the system, we observed that most variations were captured using 5 or less atoms per word; this leads to a heuristic but reasonable upper limit on the number used in the experiments. It results in an output vector of dimension 15. In the cases where the number of atoms was lower than 5, the vector was filled with zeros to meet the desired size.

The general idea behind the modelling of word level intonation for emphasis recreation is to learn which atoms should be present in the word when it is emphasised. Then, during the synthesis of F_0, a modification of intonation based on the predicted model parameters should elicit emphasis in the word. This modification consists of three steps: first, GCR model parameters are extracted from the synthetic F_0, then the atoms in the target word are removed, finally, the atoms predicted by the RF are added to the curve.

4. Experiments

4.1. Data

4.1.1. Speech material

For our experiments, we used subsets of two databases: the Wall Street Journal database [24] and the SIWIS database [25][2].

To train our random forest regressors, we used a subset of the si84 dataset from WSJ database. The 2453 sentences used were selected by comparing the output of several pitch extractors, to avoid using sentences for which the pitch would result in a poor decomposition. SSP [26], the TEMPO pitch extractor [27] and the Kaldi pitch tracker [28] were used to extract the pitch. Then, we kept only the files for which correlation was higher than 0.99 and root mean square distance (RMSE) lower than 50Hz between each pair of pitch tracker outputs.

The SIWIS database contains a set of sentences for which the speakers were asked to emphasise one predefined word. In our experiments, we used between 20 and 25 such sentences, coming from 14 speakers – speakers from subset A and C, plus speaker *B_29* – resulting in 328 sentences.

4.1.2. Feature preparation

To implement our experiments, we first needed to extract the features – GCR model parameters – from the speech. We extracted atoms with two stopping criteria: a limit of 10 atoms per second at most, and a threshold of 0.99 on the weighted correlation between F_0 and modelled F_0 between the start and end of phonation.

To scale the features to word level, we used forced aligned labels for all the data. Using the word boundaries, atoms were selected based on the position of their maximum, $t_{\max} = t_0 + (k-1)\theta$, where t_0 is the position of the impulse for the atom. 5 atoms were then selected based on their weighted correlation with the original F_0 (highest correlation first). Finally, these 5 atoms were sorted according to their position in the word, so that the final feature vector would have the shape $F_{\text{word}} = (p_1, a_1, \theta_1, p_2, a_2, \theta_2, .., p_5, a_5, \theta_5)$. That way, if there less than 5 atoms in the word, all the zeros at the end of the vector.

[2]Available at `http://bit.ly/siwisData`

Some of the contextual labels were normalized: the position of the phrase in utterance, of the word in the phrase were normalized by the length of the utterance and length of the phrase respectively. The accent and stress positions were normalized by the length of the word (number of syllables). Three stressed syllable positions in the word were kept at the maximum, and when the number of stresses was lower, the values were filled with zeros. Two stressed syllable were kept at the maximum (after trying with 3, it was found that the third accent position was not informative).

4.2. Experiment design

To be able to generate model parameters from linguistic context for emphasised words, we built two systems: one using neutral data along the emphasised data for training, and one using exclusively emphasised word samples. An HMM-based baseline was also built for comparing the performance of our approach:

- **RF1** For the case where we use only emphasised data, denoted RF1 hereafter, we first build a random forest regressor using the 303 sentences containing emphasis from the SIWIS speakers, excluding speaker *B_29*, with 15 estimators. As there are sentences with multiple emphasised words, it consists of 431 words. To try to capture speaker specific atom parameter distribution, we added 3 other estimators, using 20 words from speaker *B_29*, leaving out 5 words for testing. This was repeated to generate the 27 emphasised words that existed in the original emphasised data for this speaker, in a ten-fold cross-validation fashion.
- **RF2** The other model, denoted RF2, was first trained on the WSJ data, which amounts to approximately 60000 words, including silences and pauses. The silences and pauses, which in most of the cases do not contain atoms, may sometimes have atoms which will span on the next word. For the training, we used 25 estimators. 3 new estimators were added to the RF and trained on the 303 sentences used to train the first model set. Finally, 2 estimators were added and trained using target speaker sentences, in the same fashion as in the RF1 case.
- **Baseline** Our method is compared to an HMM-based TTS system baseline. The HMM models were trained on WSJ si84 data, and adapted to speaker *B_29* using 100 neutral sentences, using CSMAPLR [29]. Then, we adapted the models again using 20 emphasised sentences for adaptation and 5 for testing, in a ten-fold cross-validation fashion, to generate the 25 test sentences. State alignment was provided in the synthesis stage, to ease the comparison of parameters. Atoms were extracted from the synthetic version of the sentences, using the same phrase component as in the original case. This way, we could evaluate the F_0 contour at the word level, by looking only at the local components in the target word.

4.3. Evaluation

Figure 1 shows an example of reconstruction of the curve. The plot shows the original F_0, its phrase component, the F_0 generated by the baseline system, and the modified contour obtained when changing local atoms for the ones generated by the random forest regressor (in that case, the basis F_0 is from the baseline system). The coloured zone shows the boundaries of the emphasised word. The darkest continuous curve is the original F_0, extracted from natural speech. The dashed curve is the F_0 generated by the HMM-based system. The light continuous line is the reconstructed F_0 after replacing the local component of

Table 4: Average RMSE and correlation at the word level, and utterance level (log F0).

	HMM	RF1	RF2
RMSE word (log F_0)	0.14	0.12	0.11
RMSE utt (log F_0)	0.25	0.22	0.21
RMSE utt (F_0 with V/UV)	40Hz	37Hz	37Hz
Corr word (log F_0)	0.01	0.12	0.12
Corr utt (log F_0)	0.25	0.31	0.31
Corr utt (F_0 with V/UV)	0.96	0.92	0.92

the baseline system in the boundaries of the emphasised word by local components predicted by the RF1 system.

In this case, we can first observe that the light curve (RF1) is a smoothed version of the dashed curve (HMM), which is the effect of GCR decomposition, and reconstruction from atom parameters. Inside the target word, we can see that the RF curve is deviating from the HMM curve, to come closer to the original speech. Here we can see that this improvement is due to both the addition of atoms generated by the RF models and the deletion of the atoms extracted in the word for the HMM F_0 curve.

Objective measures To validate the capacity of the model to generate atoms that result in an emphasised word F_0 contour, we evaluated the performance of the prediction using standard measures: the RMSE and correlation between the reconstructed curve and the original curve. The systems were compared both at word level and at utterance level.

4.4. Results and discussion

Table 4 gives the average RMSE and correlation of the three systems at the word level for the emphasised word in each case, and at the utterance level. We look at three values for both RMSE and correlation: word level $\log F_0$ reconstruction, full sentence $\log F_0$ reconstruction, and full sentence F_0 reconstruction, taking into account only the voiced frames according to the voiced / unvoiced decision of the TEMPO pitch extractor normally used with STRAIGHT . At the word level, the curves were reconstructed by simply taking the relative position of the atom in the word, and we used a fixed length for the word, making comparisons straightforward. At the utterance level, the curve was reconstructed using the phrase component of the baseline F_0, and the atoms which were generated from the RF in both settings (RF1 and RF2). For the baseline, the synthetic F_0 was used directly.

HMM stands for the HMM-based baseline system, RF1 is the random forest regressor using only emphasised words, and RF2 is the random forest regressor when using neutral data in the training of the model.

At the word level, the result are showing quite a low correlation, especially for the baseline system. In that case, we can expect that the way parameters were extracted has an impact on the local decomposition. Because the phrase component was imposed to be the same as in the original $\log F_0$ contour, the algorithm may extract atoms in a different way to compensate the fact that this phrase component is not fitting optimally the synthetic $\log F_0$, e.g. in some cases where the contour is actually lower than the phrase component, negative atoms would be extracted, which may lead to negative correlation for some word level contours. There is no significant difference between the two other systems for this measure. The RMSE at the word level is showing similar trend, with similar results for the baseline and the RF models. The RF models show slightly lower RMSE, but with no significant difference.

When looking at the whole sentence, the baseline shows worse correlation when using the $\log F_0$, but higher correlation when calculating it only on voiced frames. On the other hand, the RMSE is slightly lower in the RF cases compared to the HMM. As explained in the discussion on figure 1, the fact that we use a parametric version of the synthetic curve along with the atoms generated by the RF models results in a smoother version, which may allow to reduce some error, hence the lower RMSE. At the same time, it can explain that correlation is a bit higher in the HMM case, because the model may smooth some patterns which should actually be modelled. One thing that should be underlined is that the HMM models have been adapted using emphasised data, and that the synthetic speech sounds generally more pronounced than before emphasis-specific adaptation. However, in the case where we did not use time aligned labels, the duration prediction output extremely slow speech, compared to the neutral model.

The fact that no – or little – difference was observed between the systems RF1 and RF2 is interesting as it suggests that using neutral data along with the emphasised data is not helping to build more efficient random forest. We would expect that using more data would help to train more robust trees, but this finding suggests, as was hinted by the observations on mutual information analysis in section 2.3, that the atoms which are present in an emphasised word, and their combination, are different from the ones that can be found in a neutral word.

5. Conclusion

In the context of emphasis recreation in TTS, we proposed an approach to recover intonation pattern of emphasis. This method, based on a generalised command-response intonation model and random forest to predict its parameters, showed similar results to the state of the art HMM-based joint modelling of intonation with other acoustic parameters, while using only small contextual factors and small amount of data. While this method is not aimed at predicting pitch for a whole utterance, it shows that it is possible to learn some patterns of intonation in the emphasised case, using a parameterisation peculiar to this intonation model. One interesting finding was that the atoms and their combinations seem to be different in the case of an emphasised word compared to a neutral case. This was demonstrated by analyses of mutual information between emphasised and neutral speech, and by experiments using two possible ways to train the random forest, one with a lot of neutral data along with emphasised data, the other with only emphasised data. Whilst our conclusions are reasonable based on objective measures, they remain to be verified using subjective listening tests.

As the GCR model is theoretically language independent, it should be possible to explore the modelling of emphasis-specific intonation in multiple language. Having multiple systems would then lead to the possibility to alter the output of synthesisers in the context of S2ST. This is one of our lines of future work.

6. Acknowledgements

This work has been conducted with the support of the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS).

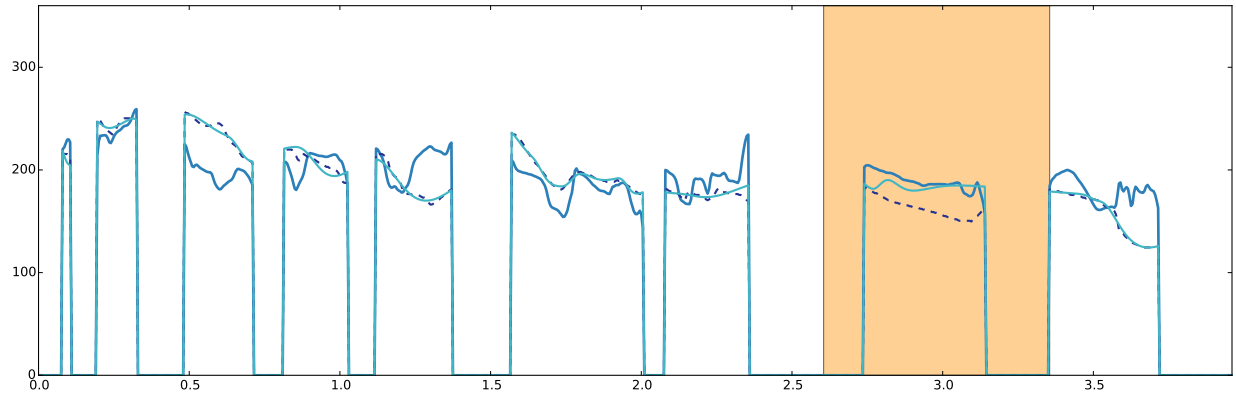

Figure 1: *Example of reconstructed F_0 contour for the sentence* "The Commission has debated the action plan for the next FIVE years." *The continuous darker blue curve is the original F_0, the dashed and darkest one is the baseline synthetic F_0, and the lightest continuous one is the proposed one.*

7. References

[1] A. Tsiartas, P. G. Georgiou, and S. S. Narayanan, "A study on the effect of prosodic emphasis transfer on overall speech translation quality," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, Canada: IEEE, 2013.

[2] M. Kurimo, W. Byrne, J. Dines, P. N. Garner, M. Gibson, Y. Guan, T. Hirsimäki, R. Karhila, S. King, H. Liang, K. Oura, L. Saheer, M. Shannon, S. Shiota, J. Tian, K. Tokuda, M. Wester, Y.-J. Wu, and J. Yamagishi, "Personalising speech-to-speech translation in the EMIME project," in *Proceedings of the ACL 2010 System Demonstrations*, Uppsala, July 2010, pp. 48–53.

[3] A. Parlikar, A. W. Black, and S. Vogel, "Improving speech synthesis of machine translation output," in *Proceedings of Interspeech*, Makuhari, Japan, September 2010, pp. 194–197.

[4] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "Intent transfer in speech-to-speech machine translation," in *Proceedings of the fourth IEEE Workshop on Spoken Language Technology*, 2012, pp. 153–158.

[5] Q. T. Do, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Preserving word-level emphasis in speech-to-speech translation using linear regression HSMMs," in *Proceedings of Interspeech*, Dresden, Germany, September 2015, pp. 3665–3669.

[6] Q. T. Do, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Improving translation of emphasis with pause prediction in speech-to-speech translation systems," 2015.

[7] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in HMM-based speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4238–4241.

[8] K. Hirose, K. Ochi, R. Mihara, H. Hashimoto, D. Saito, and N. Minematsu, "Adaptation of prosody in speech synthesis by changing command values of the generation process model of fundamental frequency," in *Proceedings of Interspeech*, Florence, August 2011, pp. 2793–2796.

[9] H. Fujisaki and S. Nagashima, "A model for the synthesis of pitch contours of connected speech," Engineering Research Institute, University of Tokyo, Tech. Rep., 1969.

[10] P.-E. Honnet and P. N. Garner, "Intonation atom based emphasis transfer," Idiap, Idiap-RR Idiap-Internal-RR-84-2015, 2015.

[11] M. Cernak and P.-E. Honnet, "An empirical model of emphatic word detection," in *Proceedings of Interspeech*, Dresden, Germany, September 2015.

[12] M. Cernak, A. Asaei, P.-E. Honnet, P. N. Garner, and H. Bourlard, "Sound pattern matching for automatic prosodic event detection," Idiap, Idiap-RR Idiap-RR-03-2016, 3 2016.

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

[14] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *ICASSP*. IEEE, 2013, pp. 7962–7966.

[15] P.-E. Honnet, B. Gerazov, and P. N. Garner, "Atom decomposition-based intonation modelling," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Brisbane, Australia: IEEE, April 2015, pp. 4744–4748.

[16] B. Gerazov, P.-E. Honnet, A. Gjoreski, and P. N. Garner, "Weighted correlation based atom decomposition intonation modelling," in *Proceedings of Interspeech*, Dresden, Germany, September 2015.

[17] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing. acoustical analysis and physiological interpretations," *Dept. for Speech, Music and Hearing, Tech. Rep*, 1981.

[18] ——, "In search for models in speech communication research," in *Proceedings of Interspeech*, Brisbane, September 2008.

[19] K. Hirose, H. Hashimoto, J. Ikeshima, and N. Minematsu, "Fundamental frequency contour reshaping in HMM-based speech synthesis and realization of prosodic focus using generation process model," in *Speech Prosody*, May 2012.

[20] H. Kameoka, J. Le Roux, and Y. Ohishi, "A statistical model of speech F0 contours," in *Proceedings ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition (SAPA)*, September 2010, pp. 43–48.

[21] H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, and K. Kashino, "Generative modeling of voice fundamental frequency contours," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1043–1052, June 2015.

[22] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *Journal of the Acoustical Society of America*, vol. 107, pp. 1697–1714, March 2000.

[23] H. Strik, "Physiological control and behaviour of the voice source in the production of prosody," Ph.D. dissertation, Dept. of Language and Speech, Univ. of Nijmegen, Nijmegen, Netherlands, October 1994.

[24] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the Workshop on Speech and Natural Language*, Stroudsburg, PA, USA, 1992, pp. 357–362.

[25] J.-P. Goldman, P.-E. Honnet, R. Clark, P. N. Garner, M. Ivanova, A. Lazaridis, H. Liang, T. Macedo, B. Pfister, M. S. Ribeiro,

E. Wehrli, and J. Yamagishi, "The SIWIS database: a multilingual speech database with acted emphasis," in *Proceedings of Interspeech*, San Francisco, USA, September 2016.

[26] P. N. Garner, M. Cernak, and P. Motlicek, "A simple continuous pitch estimation algorithm," *IEEE Signal Processing Letters*, 2012.

[27] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[28] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014, pp. 2513–2517.

[29] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.