IDIAP RESEARCH REPORT

INTONATION ATOM BASED EMPHASIS TRANSFER

Pierre-Edouard Honnet[a] Philip N. Garner

Idiap-RR-14-2016

MAY 2016

[a]Idiap Research Institute

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Intonation atom based emphasis transfer

Pierre-Edouard Honnet and Philip N. Garner

April 5, 2016

Abstract

Speech to speech translation can benefit from translation of emphasis. We propose to use an intonation model to retrieve and transfer events associated with emphasis in the intonation. This model decomposes the F_0 contour into basic intonation atoms using the matching pursuit algorithm. We investigate the role of these components in the perception of emphasis. Some of the most prominent local components are used to convert a neutral sentence to a sentence with emphasis on a specific word. The method is evaluated using parallel emphatic speech in the same language and listening tests are conducted to validate its efficacy. The results show that our intonation based approach to emphasis transfer elicits emphasis perception in neutral speech.

Index Terms: Intonation, emphasis, atom-based model, text-to-speech synthesis

1 Introduction

Speaking different languages is often a barrier for communication, that speech to speech translation (S2ST) attempts to cross. Some S2ST commercial applications have recently started to reach the mass audience, e.g. Skype™ [1]. 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 each part of the system simply pipelined into the next one. The main goal of S2ST being 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] investigated the effect of emphasis transfer on speech translation quality. In a large scale human evaluation framework, they 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[2] [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, replace untranslated words with fillers and 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.

With colleagues, we recently developed an intonation model [6, 7] that decomposes intonation into global and local components in a similar fashion as the command-response (CR) model [8]. It is a generalisation of the CR model which tries to describe the intonation contour with physiologically plausible components.

In this paper, we investigate the use of our generalised command-response (GCR) intonation model as an analysis tool for emphasis transfer in intra-lingual and cross-lingual cases.

The present work, motivated by emphasis transfer accross languages, is restricted to an intra-lingual study of emphasis transfer on natural speech. As parameter distributions from the GCR model obviously vary from one language to another, we first investigate how these parameters are related to emphasis in

[1] `http://www.skype.com`
[2] `http://www.emime.org/`

the monolingual case. However, the GCR model ought to be applicable in a cross-lingual setting, because it models intonation in a language independent way. After some observations on the links between atoms and emphasis, we hypothesise that transferring the most prominent local components of the model from an emphasised version of the word will elicit the perceived emphasis in a neutral sentence. We first extract the components of the model from acoustic features related to prosody; then, given the localisation of the emphasis and the position of the corresponding word in a neutral sentence, we transfer the local components of the model to the neutral version of the sentence.

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) [9], or more recently deep neural networks (DNNs) [10]. 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 [6, 7]. Fujisaki and colleagues have worked for several decades on a model which tries to model the underlying process of human intonation production [8, 11, 12]. 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 [13]. In a similar fashion, the CR model was used for intonation contour reshaping to add focus in the synthetic speech [14]. The CR model was also implemented as an intonation generation model using specific topology hidden Markov models [15, 16].

Anumanchipalli *et al.* [4] exploited the *tilt* model [17] to train a conversion function between vectors from input and output languages from a parallel corpus.

2.2 Generalised command-response approach

We proposed the generalised command response model as an alternative command response model characterised by an automatic parameter extraction procedure [6]. The decomposition of the contour is based on the matching pursuit algorithm with a dictionary of critically damped system impulse responses of the form of $G_{k,\theta}(t)$ (1), that happen to have the same functional form as a gamma distribution:

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

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 and by using a different global component shape, similar to (1) with higher values for θ. For more details, see [6, 7].

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 order of system, k in (1), is fixed as we assume the same order for the different impulse responses.

2.3 Application of GCR model to emphasis and 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. [18, 19]). Given parallel data including neutral and emphasised speech, we can retrieve the parameters of our model for both sentences. We hypothesise that the most prominent components in the emphasised word carry information about the emphasis; transferring these components to the corresponding word in the neutral sentence should generate artificial emphasis in an originally neutral sentence. Emphasis is expressed in different ways for different speakers, so we do not expect the mimicked emphasis to be exactly the one a speaker would naturally produce, however we expect that when modifying intonation, listeners will perceive emphasis on the target word.

Table 1: Number of atoms and additional duration (in sec.) needed on average for target word per speaker.

Spkr	Neut.	Emph.	Diff. (norm)	Diff dur.
29 (EN)	5.28	7.48	2.20 (-1.98)	0.28
26 (EN)	5.36	8.80	2.03 (2.58)	0.20
29 (FR)	5.08	7.96	2.88 (-2.56)	0.33
28 (FR)	7.20	10.56	3.36 (-1.05)	0.27
ALL	5.73	8.70	2.97 (-0.75)	0.27

3 Emphasis in atoms

3.1 Data

Emphasis analysis: Our goal is to investigate local emphasis on some words in full sentences. For that, we use a part of the multilingual SIWIS database [20][3], which consists of a parallel set of sentences: each sentence was uttered once in a neutral way, and once with specific focus on a predefined word. The speakers were told which word to emphasise before reading the focused version.

For analysis, we selected three speakers numbered *26*, *28* and *29*. For speaker *29* (female), we used both English and French data, for speaker *26* (male) English data and for speaker *28* (male) French data. These speakers were selected because the vocoder used (STRAIGHT [21]) worked well for them. For each language and each speaker, we use 25 neutral sentences and the 25 same sentences with emphasis on a word. Thus, 100 neutral sentences are compared with their emphatic versions.

Emphasis transfer: The data used for the emphasis transfer experiment is a subset of the dataset described above: speaker *29* was selected, and the evalution was carried out only on the English sentences, to ease subjective listening tests.

3.2 Comparing neutral and emphatic sentences

For each sentence in our dataset, we perform a decomposition of the F_0 using the GCR model, and then compare some general statistics on the parameters. The order of the impulse responses is fixed to $k = 6$, following the results found in our previous work [7], the dictionary is composed using atoms with $\theta \in \{0.01, 0.015, .., 0.05\}$. Our hypothesis is that the statistics on parameters will differ between the neutral and emphasised case.

We first look at the number of atoms needed to model the local behaviour of F_0 in the emphasised word. We do not investigate duration modifications in this work. However, to compare the number of commands in the neutral and focused case, we measure the duration of the word under investigation for each sentence in table 1. The average difference between the duration of the emphasised word and the neutral word is calculated and given with the average number of atoms and their difference (emphasised - neutral) in the two contexts for each speaker. The difference between the number of atoms required for emphasised and neutral cases is also given when normalised over time in brackets in the 4^{th} column.

As we might expect, more atoms are needed to model the target word in the emphasised case. We might think that one of the reasons for this is the fact that the words have a longer duration, but looking at the difference in number of atoms normalised over the duration of the words (in brackets in the 4^{th} column), we can see that in average, there are fewer atoms per second in the emphasised version of the word. This is interesting as it shows that the way the atoms are distributed in the emphatic word is not only related to the duration of the word, as compared to the neutral case.

By comparison, the regions outside the target word typically have 30 atoms, and require just 3 more on average in the emphatic case. The ratio of numbers of atoms between emphatic and neutral is 1.1 ± 0.04 on average, which can be explained by a slightly slower speaking rate, used for increasing the emphasis on the target word (for duration, the ratio is 1.19 ± 0.02).

We also looked at the mutual information between atom parameters and some linguistic features to observe the differences between emphasised and neutral words in the same context, at the syllable level. The investigated features were accent, stress and emphasis. The mutual information between labels L and model features F_i was calculated as follows:

[3]The current version of the database is available at `http://www.unige.ch/lettres/linguistique/recherche/latl/siwis/database/`

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**

$$I(L, F_i) = \sum_{l} \sum_{f \in F_i} p(l, f) \log_2 \left(\frac{p(l, f)}{p(l)p(f)} \right) \tag{2}$$

where $p(l, f)$ is a joint probability of L and F_i, and $p(l)$ and $p(f)$ are their respective marginal probabilities. These probabilities are calculated for each class according to the following quantisation: between 0 and 10 for position (relative position in the syllable), and between 0 and 9 for amplitude. The labels l are binary. We normalise the mutual information with the entropy of the contextual labels, defined as:

$$H(L) = -\sum_{l} p(l) \log_2(p(l)) \tag{3}$$

Table 2 shows the values for $\frac{I(L,F_i)}{H(L)}$. These results were obtained on a bigger set from the SIWIS database (about 300 sentences in each case, emphasised and neutral).

In each case, we give the mutual information between accent, stress, emphasis or a combination of them and amplitude, position, or number of atoms in the syllable in two cases: the neutral (left) case and the emphasised case (right, the syllable belongs to an emphasised word). These results reinforce the previous observation, as the number of atoms per syllable seems a dominant feature for emphasis. When looking at the mutual information between the number of atoms at the word level and the emphasis, we found a mutual information of 28.3 in the neutral case, against 46.8 in the emphasised case.

When looking at the amplitude of the atoms, we found that in some cases there were some atoms with slightly higher amplitude in the emphasised case, but with no significant differences. The distribution of the extracted atom θ did not show any significant difference between neutral and emphatic versions of the target words. In the same line, from a mutual information point of view, amplitude and θ do not seem to be relevant to discriminate emphasis. This is interesting as it indicates that emphasis would manifest itself as more atoms rather than higher amplitude versions of those from the neutral case. The hypothesis that the atom parameters differ from one version to another is refuted, however we observe a different way to decompose the intonation – with more components.

4 Experiments

4.1 Emphasis transfer

Based on the observations made in 3.2, we know that more atoms are representative of emphasis. However, our observations do not allow us to conclude on which atoms are responsible for emphasis. Our natural hypothesis is that the atoms which have the highest absolute amplitude are the most important in the expression of emphasis through intonation. Then, we expect that by adding these atoms with the highest absolute amplitude – they can be positive or negative – at the appropriate position in the neutral sentence, emphasis will be perceived better and stronger on the target word. To assess the emphasis perception, we run a subjective listening test. If the emphasis is rated higher and on the target word when transferring atoms, it will show that these atoms play an important role in the emphasis delivery.

We aim at generating artificial emphasis on a target word in a neutral sentence by altering only intonation. To do so, we extract the model parameters for the full sentence in each case, and given time alignment and the knowledge of which word are emphasised, we identify the atoms in the target word. As the two versions of the sentence have different durations, for the target word and the other words, we calculate the relative position of the atoms in each syllable and transfer them to the corresponding syllable in the neutral sentence target word.

To show that most prominent atoms are important for emphasis perception, we select only the atoms with the highest absolute amplitude and transfer them to the approximate position in the neutral sentence. In that particular context, it is easy to find the corresponding position, because the words are the

same and thus have the same number of syllables; moreover we can assume that their relative durations are extended in a similar fashion.

Initial experiments showed that adding preceding syllable atoms did not bring perceivable difference, therefore only the main atoms in the target word were transferred. It was found empirically that transferring more than 3 components did not improve the perception of emphasis, thus only 3 atoms – or fewer in the case where there were fewer atoms in the word – were given to the neutral sentence. It is also in line with the average additional number of atoms needed to model the target word found in section 3.2. In particular, for the 20 sentences selected for this speaker, we found that 2.4 more atoms were needed on average in the emphatic case.

4.2 Subjective evaluation

A subjective listening test was conducted to evaluate the validity of our approach. The listeners were asked to listen to the samples in a random order and identify which word sounded the most emphasised, and for this word give a level of emphasis with a 3-level choice: *clear*, *moderate* or *slight* emphasis. Each subject had to listen to 3 versions of 20 sentences, S1–S20, for a total of 60 audio samples. One version consisted of the original neutral sentence, another one the original sentence with emphasis, and the last one the neutral sentence with artificial emphasis. An example for each level of emphasis was given in the instructions, to understand how to rate the degree of emphasis. The listeners always had to identify a "most emphasised" word in order to control that emphasis transfer had an effect compared to the neutral sentences. We expected listeners to rate the neutral sentences as *slightly* emphasised, the original emphatic version as *clearly* or *moderately* emphasised, and the artifically emphasised version closer to the emphatic version, as the aim is to increase the impression of emphasis on the target word.

30 subjects participated in the test, with a high majority of non native fluent English speakers, most of them being in the age range 26-35.

5 Results and discussion

5.1 Results

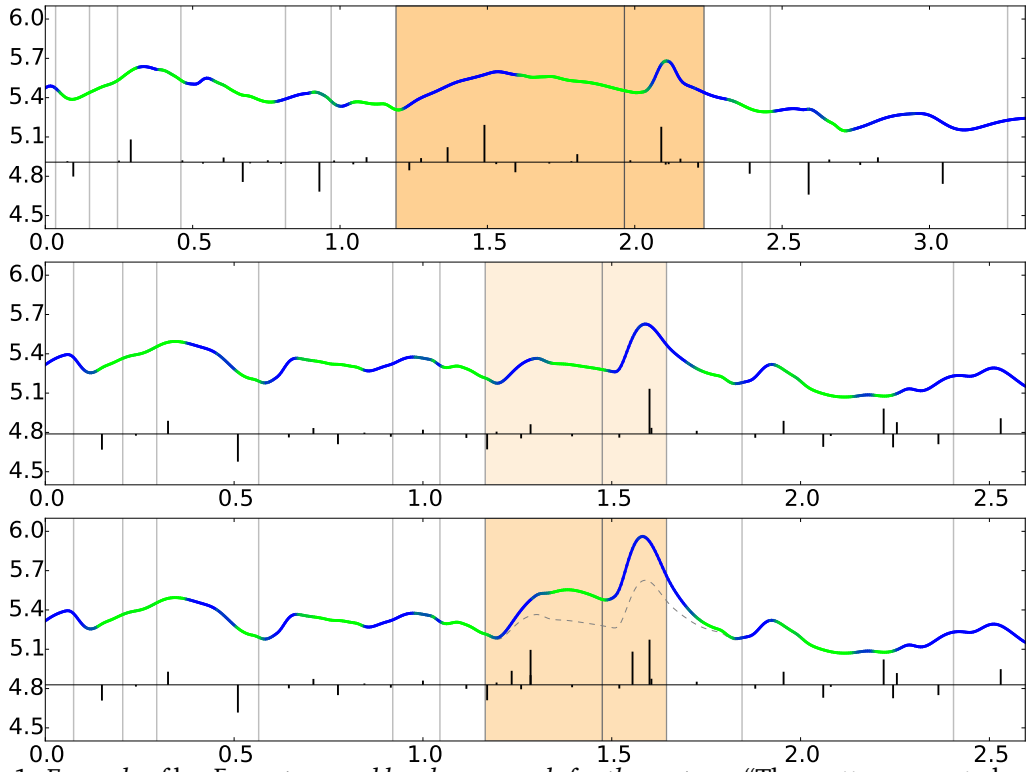

Figure 1: *Example of* log F_0 *contour and local commands for the sentence* "The matter seems to be somewhat confused.". *Top panel: sentence with emphasis on the word* "somewhat". *Middle: neutral sentence. Bottom: neutral sentence with transferred emphasis on the word* "somewhat".

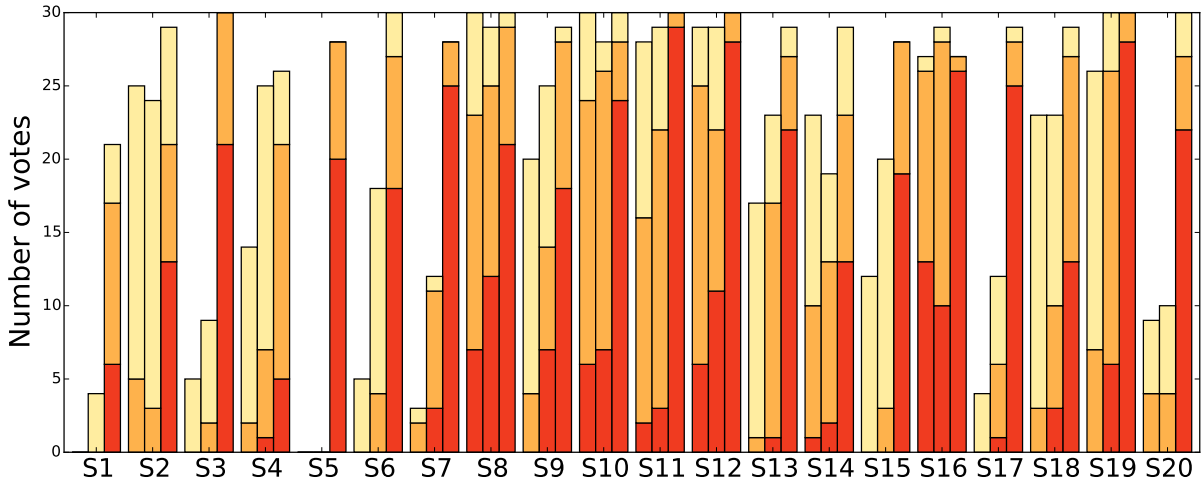

Figure 2: *Subjective listening test results. Red = clear, orange = moderate, yellow = slight emphasis. Missing bar means 0 vote for the target word.*

Figure 1 shows an example of transfer for a sentence with the two contours of the same sentence in the two different contexts, and the resulting contour (*S6* in the results). For the $\log F_0$ curves, the green indicates a high probability of voicing, while the blue indicates a high probability of being in an unvoiced region. The syllable boundaries are displayed, with the lightly coloured region being the target word, *"somewhat"*. In the bottom panel, we can see the original neutral contour in dashed grey, while the modified curve presents an increased F_0. The atom commands are displayed in black, and we can see that 3 components were added to the neutral sentence.

Figure 2 shows the number of people identifying the target word as most emphasised for each sentence. For each triplet of bars, the most left one corresponds to the neutral version of the sentence, the middle one is the neutral with emphasis transfer and the most right the original emphasised version. The height of the full bars corresponds to the number of votes for the target word independently of the level of the emphasis. The different colours account for the level of emphasis that the voters chose when they chose the target word. The darkest (bottom) colour stands for *clear* emphasis, medium for *moderate* emphasis and the lightest one (top) for *slight* emphasis.

We observe 2 main trends in the results:

- In 8 cases – *S1*, *S3*, *S4*, *S6*, *S7*, *S17*, *20* – the number of people perceiving the target word as emphasised increased. For 3 of these cases, a majority of people voted for the target word when intonation was modified. For the other 5 cases, the perception of emphasis increased significantly when modifying the intonation, but did not reach the majority of votes. These 8 cases showed that the emphasis is consistently shifted towards the target word, with a higher level of emphasis.

- In 11 other cases out of the 20, the majority of the listeners voted for the target word in the neutral case, even though the speaker did not have any particular instructions. For 4 of these cases, adding atoms decreased the number of votes for the target word, however in all these cases, the number of subjects choosing a *clear* emphasis increased, and the number of *moderate* emphasis also increased. In 2 cases the total number stayed the same, but there was an increase in the number of *clear*, and in *moderate* emphasis. In the 5 other cases, the total number always increased and the level of emphasis was also rated higher. These 11 cases showed that when the emphasis is already perceived on the target word, its strength is increased when adding emphasis atoms.

- In the last case (*S5*), adding local components from the emphasised word intonation was not enough to make the perception of emphasis change for the listeners, the target word being a non content word. Most of the listeners kept the main content word as most emphasised.

5.2 Discussion

The global trend in the results confirms the hypothesis: transferring local components from an emphasised word to a neutral sentence increases the impression of emphasis in the target word in most of the cases. We can also see that the way emphasis is perceived – in other words how strong the emphasis is –

is affected by adding local positive or negative components. The modification of the resulting intonation contour seems to increase the strength of the emphasis.

In some cases, emphasis was not perceived on the target word mainly because of the *reset* at the beginning of the sentence – sentences start with a raising intonation before gradually decreasing. It may have been confusing for the listeners – and this was a feedback from some of the listeners – to choose between a slightly emphasised word in the middle of an utterance and the natural higher pitch that occurs at the start of speech.

We cannot expect the intonation alone to help the listeners to perfectly perceive the emphasis on the target words, however the results indicate that it consistently improves the perception of the emphasis and its strength.

We should also mention the fact that this work only focussed on a particular type of emphasis, where the speakers were asked to emphasised a specific word with no further instructions. It is obviously different from the emphasis that would occur naturally when speaking, from contrastive emphasis, or from the emphasis used when conveying new information for instance.

6 Conclusion

We presented an application of the generalised command-response model for emphasis transfer. Some analyses on French and English data indicated that modelling intonation with the GCR model in the emphasised word was requiring more elements than for a neutral word, but comparitively fewer elements for the same duration. Our experiments and listening tests 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. In most cases, the number of people correctly identifying the target word when modifying intonation increased.

7 Acknowledgements

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

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] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] F. Tamburini, "Automatic prominence identification and prosodic typology," in *Proceedings of Interspeech*, 2005, pp. 1813–1816.

[20] 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*, September 2014.

[21] 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.