# Investigation into Transferability of Duration of Emphasised Words from Original Expression to Spoken Translation

梁晖 *(Hui LIANG)*

Speech Processing Group, Computer Engineering & Networks Lab, ETH Zürich, Switzerland

TIK-Report No.359, ETH Zürich, November 2014

## Abstract

To achieve an expressive, personalised speech-to-speech translator, paralinguistic information need be transferred from input natural speech into output synthesised speech, in particular, a user's focuses carried by emphasised spoken words in input speech. In order to examine if the duration of emphasised words in a source language is usable for predicting that of their emphasised spoken translations, this paper presents results of analysis of sentence-level and word-level duration extracted from parallel English and German speech utterances containing emphasis. Firstly, it was confirmed to be reasonable to consider emphasised words did not affect duration of their nearby neutral parts. Secondly, the realisation of emphasis in terms of word duration was speaker-dependent to an extent and language-independent for most speakers. Thirdly, though the duration of an emphasised word in a source language by itself was not sufficient for a definite prediction of the duration of its spoken translation in a target language, collectively, duration of emphasised words in the source language could still portray a speaker's own style of emphasising words in the target language.

**Index Terms**: emphasis, duration, cross-language transferability

## 1. Introduction

It is no longer a major challenge to generate sufficiently natural-sounding, highly intelligible speech in the neutral reading style by the state-of-the-art technology of statistical parametric text-to-speech synthesis [1, 2]. Recently, speech scientists have been much interested in the realisation of expressive prosody in synthesised speech within this statistical parametric framework, e.g. the realisation of emotions (happiness, sadness, anger, etc) [3, 4]. Apart from that, another interesting type of expressive prosody is emphasis, which can highlight the focus of an utterance and is a very common phenomenon in spontaneous speech.

The context of a conversation and intentions/moods of a speaker determine when and how to realise emphases in an utterance. For example, any word in the sentence "*I didn't take the test yesterday*" may be emphasised in various manners as per the subtext of its speaker. It is basically insurmountable for state-of-the-art front-ends to predict when and how to synthesise an emphasis that carries a special meaning or purpose merely based on plain text input to a speech synthesiser, although it is achievable to make manually specified spoken words in an utterance emphatic in a certain way [5]. In contrast, realising emphases appears to be easier in the context of a recent hot research topic, speech-to-speech translation: since output speech is meant to be identical with input speech in meaning, input prosody should be suitable to be considered a guide for the realisation of output prosody. For example, output speech should sound like "*ich habe gestern die* ***Prüüüüfung*** *nicht abgelegt*" when input speech is "*I didn't take the* ***teeeest*** *yesterday*" (repetition shows lengthening). It is reasonable to assume that prosodic information from input speech can help to realise emphases in output speech and therefore to improve speech-to-speech translation systems in terms of naturalness of synthesis [6].

In the above proposition, prosodic cues are to be transferred from a source to a target language. Research has recently been conducted on the transfer of intonation across languages for conveying emphases to a target language [7]. The approach was analogous to the common solution to voice conversion based on parallel speech data and modelling joint probability densities of source and target feature vectors by Gaussian mixture model. However, whether the prosodic cues were applicable to the target language was not explicitly investigated. Naturally, it would be of great interest to examine whether emphases are realised differently in different languages, in other words, whether their prosodic cues are transferable across languages.

It is widely agreed that prosodic features relevant to emphasis include duration, pitch and intensity [8, 9, 10]. Duration was the focus of this research work. English and German, which are stress-timed [11], were investigated in the following analyses. Results of the analyses of duration were expected to provide inspirational insights into the *cross-language* transferability of duration of emphasis – owing to this goal, *intra-lingual* correspondence amongst the words within an utterance, e.g. whether duration of the neutral neighbours of an emphasised word were affected more greatly than those of farther neutral words, was not touched in this work.

## 2. Preparations for Analysis of Duration

### 2.1. Specification of Speech Data

Speech recordings in the voices of the four bilingual (German and English) speakers (sp1, sp2, sp3 and sp4) in a pilot multilingual corpus recorded at the University of Geneva [12] were used for analysis of duration. These speech recordings had been grouped into quadruples. The four utterances in a quadruple $({}^{i}U^{\text{deu}}_{\text{neu}}, {}^{i}U^{\text{deu}}_{\text{emp}}, {}^{i}U^{\text{eng}}_{\text{neu}}, {}^{i}U^{\text{eng}}_{\text{emp}})$, where $i$ is a sentence number, were parallel data in the sense as an exemplar below shows:

- ${}^{i}U^{\text{deu}}_{\text{neu}}$: Ich habe gestern die Prüfung nicht abgelegt.
- ${}^{i}U^{\text{deu}}_{\text{emp}}$: Ich habe gestern die **Prüfung** nicht abgelegt.
- ${}^{i}U^{\text{eng}}_{\text{neu}}$: I didn't take the test yesterday.
- ${}^{i}U^{\text{eng}}_{\text{emp}}$: I didn't take the **test** yesterday.

where the superscripts $^{\text{deu}}$ and $^{\text{eng}}$ denote *German* and *English* respectively, and the subscripts $_{\text{neu}}$ and $_{\text{emp}}$ denote *neutral* and *emphasised* respectively. During the speech recording session, speakers were presented with printed sentences. When a sentence contained 1–3 words (mostly only one word) printed in capitals, the speakers were requested to read the whole sentence and to emphasise these 1–3 words in their own natural ways.

In summary, the four utterances in a quadruple have the same meaning, and the transcription of $\{{}^{i}U_{\text{emp}}\}$ is identical to that of $\{{}^{i}U_{\text{neu}}\}$. Hence, it makes sense to compare $\{{}^{i}U^{\text{deu}}\}$ and $\{{}^{i}U^{\text{eng}}\}$, and it is apparent that ${}^{i}U_{\text{neu}}$ provides an excellent reference point for measuring the degree of emphasis in ${}^{i}U_{\text{emp}}$.

### 2.2. Normalisation of Phone Duration

First of all, automatic segmentation (forced alignment) was carried out such that durations of phones in every utterance could be obtained. Because the four speakers spoke at different rates, their original speaker-specific and possibly language-dependent phone durations were normalised according to Eq. (1):

$$\hat{d}_{\text{phone}} = \frac{d_{\text{phone}}}{\mathcal{D}^{\text{SLD}}_{\text{all}}} \tag{1}$$

where $d_{\text{phone}}$ is the original duration of a phone obtained by automatic segmentation, $\hat{d}_{\text{phone}}$ is the normalised duration of this phone, $\mathcal{D}^{\text{SLD}}_{\text{all}}$ denotes the mean of durations of all the phones (no silence and pause) in German or in English uttered by a particular speaker (i.e. $\mathcal{D}^{\text{SLD}}_{\text{all}}$ is both *s*peaker-specific and *l*anguage-*d*ependent). Every subsequent analysis was conducted on the basis of the entire set $\{\hat{d}_{\text{phone}}\}$, and no segments of silence and pauses were taken into consideration.

### 2.3. Two-Sample Kolmogorov-Smirnov Test

The two-sample Kolmogorov-Smirnov (K-S) test [13, 14] is a non-parametric hypothesis test that returns a decision as to whether two independent random samples are drawn from the same underlying one-dimensional continuous population by evaluating the difference between the empirical cumulative distribution functions (cdfs) of the two samples. The *two-sided* two-sample K-S test, which was utilised in this research work, employs the maximum *absolute* difference between the two empirical cdfs as the test statistic. The null hypothesis $H_0$ of a two-sample K-S test is that the two samples are from the same distribution. The significance level $\alpha$ is typically 0.05.

## 3. Analyses of Duration Data

In this section, when a set of data is presented in the form of a box plot, a median is indicated by a solid bar across a box that shows quartiles. Whiskers extend up to 1.5 times an interquartile range and plus signs represent “outliers” that are beyond such an extended range.

### 3.1. Analysis at the Sentence Level

First of all, an analysis of duration was conducted at the sentence level: all the phones of the emphasised part of an utterance ${}^{i}U_{\text{emp}}$ were collectively regarded as $\mathcal{P}^{i}_{\text{emp}}$, while all other phones in this utterance were collectively regarded as $\mathcal{P}^{i}_{\text{neu}}$. For the sake of a sensible and fair analysis, the corresponding utterance ${}^{i}U_{\text{neu}}$ was considered a reference point so that the measurements calculated as per Eq. (2) were actually analysed:

$${}^{i}\hat{r}^{\text{[lang]\_[type]}}_{\text{sentence}} = \frac{\sum_{\mathcal{P}^{i}_{\text{[type]}} \text{ of } {}^{i}U^{\text{[lang]}}_{\text{emp}}} \hat{d}_{\text{phone}}}{\sum_{\text{neutral version in } {}^{i}U^{\text{[lang]}}_{\text{neu}} \text{ for } \mathcal{P}^{i}_{\text{[type]}}} \hat{d}_{\text{phone}}} \tag{2}$$

where [lang] is *deu* or *eng*, and [type] is *neu* or *emp*.

Figure 1 presents an overview of $\{{}^{i}\hat{r}^{\text{[lang]\_[type]}}_{\text{sentence}}\}$, i.e. relative normalised durations at the sentence level. This figure demonstrates the common knowledge that people tend to lengthen words when emphasising them. In addition, this figure shows that lengthening emphasised words only marginally affected the durations of neutral parts of $\{{}^{i}U_{\text{emp}}\}$, no matter whether the speakers spoke English or German (see *deu_neu* and *eng_neu*). A two-sample K-S test was performed to compare pairwise the similarity of the four cases in Figure 1. The result is listed in Table 1.

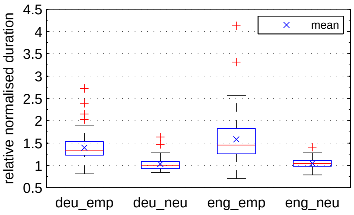

Figure 1: Overview of $\{{}^{i}\hat{r}^{\text{[lang]\_[type]}}_{\text{sentence}}\}$

Table 1: Result of two-sample K-S test on $\{{}^{i}\hat{r}^{\text{[lang]\_[type]}}_{\text{sentence}}\}$

| | deu_emp | deu_neu | eng_emp | eng_neu |
|---|---|---|---|---|
| deu_emp | – | $<10^{-12}$ | 0.0566 | $<10^{-15}$ |
| deu_neu | yes | – | $<10^{-15}$ | 0.2435 |
| eng_emp | no | yes | – | $<10^{-15}$ |
| eng_neu | yes | no | yes | – |

Upper triangle: $p$ value ($\alpha = 0.05$)
Lower triangle: answer to “Reject $H_0$?”

Table 1 shows two null hypotheses could not be rejected at $\alpha = 0.05$: (1) $\{{}^{i}\hat{r}^{\text{deu\_neu}}_{\text{sentence}}\}$ and $\{{}^{i}\hat{r}^{\text{eng\_neu}}_{\text{sentence}}\}$ were from the same distribution; (2) $\{{}^{i}\hat{r}^{\text{deu\_emp}}_{\text{sentence}}\}$ and $\{{}^{i}\hat{r}^{\text{eng\_emp}}_{\text{sentence}}\}$ were from the same distribution. The first one confirms that the usual durations of the neutral parts of the utterances containing emphasis were maintained in English and German consistently. The $p$ value of the second one was close to $\alpha$. Hence a closer look will be taken later.

### 3.2. Analysis of Neutral Spoken Words

The analysis at the sentence level in section 3.1 provides a general picture of language-dependent, speaker-independent relative normalised sentence duration. Since the neutral part of an utterance ${}^{i}U_{\text{emp}}$ contained more than one word, the word-level measurements shown in Eq. (3) were also analysed:

$${}^{i}\hat{r}^{\text{[lang]\_neu}}_{\text{word}} = \frac{\sum_{\text{all phones of neutral spoken word } \mathcal{W} \text{ in } {}^{i}U^{\text{[lang]}}_{\text{emp}}} \hat{d}_{\text{phone}}}{\sum_{\text{all phones of neutral version of } \mathcal{W} \text{ in } {}^{i}U^{\text{[lang]}}_{\text{neu}}} \hat{d}_{\text{phone}}} \tag{3}$$

where [lang] is *deu* or *eng*.

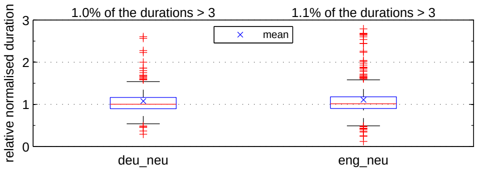

Figure 2: Box plot of speaker-independent $\{{}^{i}\hat{r}^{\text{[lang]\_neu}}_{\text{word}}\}$

Figure 2 presents an overview of $\{{}^{i}\hat{r}^{\text{[lang]\_neu}}_{\text{word}}\}$, i.e. relative normalised durations of neutral spoken words in $\{{}^{i}U_{\text{emp}}\}$. This figure shows similar concentration of relative normalised durations around 1.0 as Figure 1 shows. Lengthening emphasised words did not highly affect the durations of neutral spoken words in $\{{}^{i}U_{\text{emp}}\}$, no matter whether the speakers spoke

German or English. The main difference, not surprisingly, is that the variation at the word level was larger than that at the sentence level. A two-sample K-S test showed the null hypothesis that speaker-independent $\{{}^{i}\hat{r}^{\text{deu\_neu}}_{\text{word}}\}$ and $\{{}^{i}\hat{r}^{\text{eng\_neu}}_{\text{word}}\}$ were from the same distribution could not be rejected at $\alpha = 0.05$ ($p = 0.6892$). This suggests that language was not an influential factor in durations of neutral spoken words in $\{{}^{i}U_{\text{emp}}\}$.

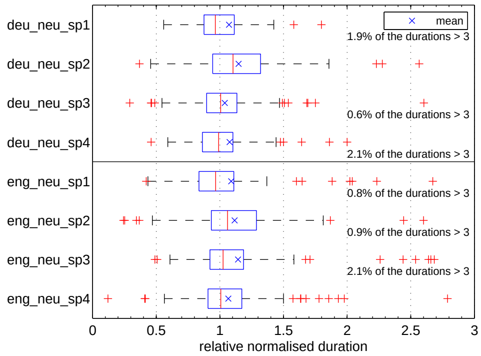

Figure 3: Box plot of speaker-dependent $\{{}^{i}\hat{r}^{\text{[lang]\_neu}}_{\text{word}}\}$

Speaker-dependent $\{{}^{i}\hat{r}^{\text{[lang]\_neu}}_{\text{word}}\}$ were then investigated, as shown in Figure 3. Speaker-dependent $\{{}^{i}\hat{r}^{\text{[lang]\_neu}}_{\text{word}}\}$ apparently concentrated around 1.0 in a fairly consistent fashion. A two-sample K-S test was performed to examine the observations from Figure 3 (see results in Tables 2 and 3).

Table 2: Result of *cross-speaker* two-sample K-S test on speaker-dependent $\{{}^{i}\hat{r}^{\text{[lang]\_neu}}_{\text{word}}\}$

| deu_neu | sp1 | sp2 | sp3 | sp4 |
|---|---|---|---|---|
| sp1 | – | 0.00017 | 0.2316 | 0.7403 |
| sp2 | yes | – | 0.00077 | 0.0011 |
| sp3 | no | yes | – | 0.6832 |
| sp4 | no | yes | no | – |
| eng_neu | sp1 | sp2 | sp3 | sp4 |
| sp1 | – | 0.0034 | 0.0276 | 0.1395 |
| sp2 | yes | – | 0.4424 | 0.2781 |
| sp3 | yes | no | – | 0.8182 |
| sp4 | no | no | no | – |

Upper triangle: $p$ value ($\alpha = 0.05$)
Lower triangle: answer to "Reject $H_0$?"

Table 2 shows the result of comparing $\{{}^{i}\hat{r}^{\text{[lang]\_neu}}_{\text{word}}\}$ within a language amongst the four speakers. Concerning the German language, speaker sp2 was an exception. As for the English language, speaker sp1 was a partial exception. This indicates the majority of the speakers had sufficiently similar behaviour so that the speaker differences were not an influential factor in duration of neutral spoken words in $\{{}^{i}U_{\text{emp}}\}$. Table 3 shows the result of comparing $\{{}^{i}\hat{r}^{\text{deu\_neu}}_{\text{word}}\}$ with $\{{}^{i}\hat{r}^{\text{eng\_neu}}_{\text{word}}\}$ for each individual speaker. Since that the two sets from every speaker obeyed the same distribution could not be rejected at $\alpha = 0.05$, language was confirmed not to be an influential factor in durations of neutral spoken words in $\{{}^{i}U_{\text{emp}}\}$.

Table 3: Result of *cross-language* two-sample K-S test on speaker-dependent $\{{}^{i}\hat{r}^{\text{[lang]\_neu}}_{\text{word}}\}$, $\alpha = 0.05$

| | sp1 | sp2 | sp3 | sp4 |
|---|---|---|---|---|
| Reject $H_0$? | no | no | no | no |
| $p$ value | 0.7888 | 0.6238 | 0.2525 | 0.4316 |

### 3.3. Analysis of Emphasised Words

The measurements analysed in this subsection were calculated as per Eq. (4), which was analogous to that in section 3.2:

$$ {}^{i}\hat{r}^{\text{[lang]\_emp}}_{\text{word}} = \frac{\sum_{\text{all phones of emphasised word } \mathcal{W}' \text{ in } {}^{i}U^{\text{[lang]}}_{\text{emp}}} \hat{d}_{\text{phone}}}{\sum_{\text{all phones of neutral version of } \mathcal{W}' \text{ in } {}^{i}U^{\text{[lang]}}_{\text{neu}}} \hat{d}_{\text{phone}}} \qquad (4) $$

where [lang] is *deu* or *eng*.

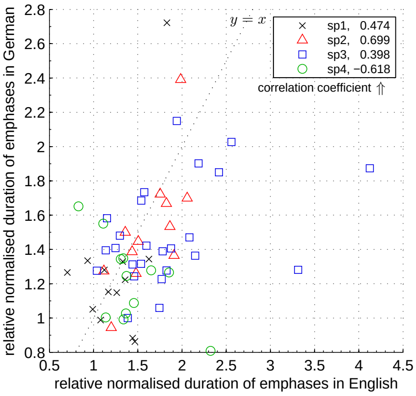

Figure 4: Scatter plot of $\{({}^{i}\hat{r}^{\text{eng\_emp}}_{\text{word}}, {}^{i}\hat{r}^{\text{deu\_emp}}_{\text{word}})\}$

Figure 4 shows speaker-specific $\{({}^{i}\hat{r}^{\text{eng\_emp}}_{\text{word}}, {}^{i}\hat{r}^{\text{deu\_emp}}_{\text{word}})\}$ in a Cartesian coordinate plane. This scatter plot reflects correspondence between emphasised words in ${}^{i}U^{\text{eng}}_{\text{emp}}$ and ${}^{i}U^{\text{deu}}_{\text{emp}}$, which were in the same quadruple of utterance. Since $\{{}^{i}U^{\text{eng}}_{\text{emp}}\}$ and $\{{}^{i}U^{\text{deu}}_{\text{emp}}\}$ were parallel data, the fact that most data points are near the line $y = x$ indicates that the speakers lengthened emphasised words in the two languages similarly to a certain extent. However, only a few data points are extremely close to the line $y = x$ and the correlation coefficients are rather undesirable. Naturally enough, the speakers could not control precisely the time for emphasising words when reading ${}^{i}U^{\text{eng}}_{\text{emp}}$ and ${}^{i}U^{\text{deu}}_{\text{emp}}$, which resulted in a great deal of randomness. Because of the inevitable randomness and the fact that a speaker read each sentence only once, it would make more sense to investigate $\{{}^{i}U^{\text{eng}}_{\text{emp}}\}$ and $\{{}^{i}U^{\text{deu}}_{\text{emp}}\}$ collectively, rather than to look into the utterances quadruple by quadruple.

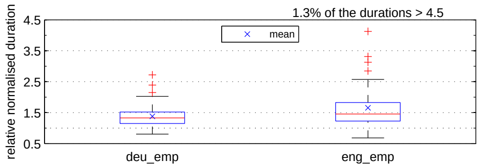

Figure 5: Box plot of $\{{}^{i}\hat{r}^{\text{[lang]\_emp}}_{\text{word}}\}$

Figure 5 presents an overview of $\{{}^{i}\hat{r}^{\text{[lang]\_emp}}_{\text{word}}\}$, i.e. relative normalised durations of emphasised words in $\{{}^{i}U_{\text{emp}}\}$. According to Figure 5, emphasised words in English were lengthened to a greater extent than those in German. A follow-up two-sample K-S test showed that the null hypothesis that speaker-independent $\{{}^{i}\hat{r}^{\text{deu\_emp}}_{\text{word}}\}$ and $\{{}^{i}\hat{r}^{\text{eng\_emp}}_{\text{word}}\}$ were from the same distribution could be rejected at $\alpha = 0.05$ ($p = 0.0229$). Consequently, language seems to have affected the manner of emphasis in terms of word duration.

Speaker-dependent $\{{}^{i}\hat{r}^{\text{[lang]\_emp}}_{\text{word}}\}$ are plotted in Figure 6 for further analysis. Speaker-dependent $\{{}^{i}\hat{r}^{\text{[lang]\_emp}}_{\text{word}}\}$ are much less consistent than speaker-dependent $\{{}^{i}\hat{r}^{\text{[lang]\_neu}}_{\text{word}}\}$ (see the contrast

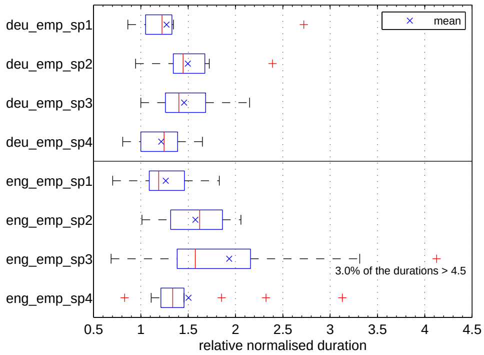

Figure 6: Box plot of speaker-dependent $\{{}^{i}\hat{r}^{\text{[lang]\_emp}}_{\text{word}}\}$

between Figures 6 and 3). A two-sample K-S test was performed again to examine the observations from Figure 6 and the results are listed in Tables 4 and 5.

Table 4: Result of *cross-speaker* two-sample K-S test on speaker-dependent $\{{}^{i}\hat{r}^{\text{[lang]\_emp}}_{\text{word}}\}$

| deu_emp | sp1 | sp2 | sp3 | sp4 |
|---|---|---|---|---|
| sp1 | – | 0.0012 | 0.0114 | 0.4443 |
| sp2 | yes | – | 0.8047 | 0.0280 |
| sp3 | yes | no | – | 0.0874 |
| sp4 | no | yes | no | – |
| **eng_emp** | **sp1** | **sp2** | **sp3** | **sp4** |
| sp1 | – | 0.0943 | 0.0144 | 0.3792 |
| sp2 | no | – | 0.3964 | 0.2672 |
| sp3 | yes | no | – | 0.0085 |
| sp4 | no | no | yes | – |

Upper triangle: $p$ value ($\alpha = 0.05$)
Lower triangle: answer to "Reject $H_0$?"

Table 4 shows the result of comparing $\{{}^{i}\hat{r}^{\text{[lang]\_emp}}_{\text{word}}\}$ within a language amongst the four speakers. Concerning the German language, both sp1 and sp2 were rather more distinctive. As for the English language, speaker sp3 was an exception. Hence, the speaker differences had a more noticeable impact on durations of emphasised words in $\{{}^{i}U_{\text{emp}}\}$. Table 5 shows the result of comparing $\{{}^{i}\hat{r}^{\text{deu\_emp}}_{\text{word}}\}$ with $\{{}^{i}\hat{r}^{\text{eng\_emp}}_{\text{word}}\}$ for each individual bilingual speaker. Only the null hypothesis that the two sets of speaker sp3 obeyed the same distribution could be rejected at $\alpha = 0.05$, but the $p$ value was extremely close to $\alpha$. This indicates that for most of the bilingual speakers language was not an influential factor in durations of emphasised words in $\{{}^{i}U_{\text{emp}}\}$.

Table 5: Result of *cross-language* two-sample K-S test on speaker-dependent $\{{}^{i}\hat{r}^{\text{[lang]\_emp}}_{\text{word}}\}$, $\alpha = 0.05$

| | sp1 | sp2 | sp3 | sp4 |
|---|---|---|---|---|
| Reject $H_0$? | no | no | yes | no |
| $p$ value | 0.6324 | 0.1311 | 0.04998 | 0.1971 |

## 4. Discussions

The observations in section 3 basically suggest that durations of neutral parts and neutral words of utterances containing emphasis were only slightly affected by nearby emphasised spoken words, and that this occurred consistently in English and in German for every bilingual speaker. Hence it is reasonable not to take into account transferring duration from the neutral part of input natural speech to that of output synthesised speech, when one wants to build an expressive speech-to-speech translator between German and English.

These observations also suggest that it is difficult to predict the exact duration of the spoken translation in German according to that of an original emphasis in English, and vice versa, although the duration of an emphasis in a source language can provide a vague cue. In spite of that, collectively speaking, emphases in English and German were not produced differently in terms of duration at the sentence level at the significance level of 0.05. At the word level, the observations in section 3.3 suggest that speakers should be treated individually, and that most of the time speech in a source language can be collected from a user to examine whether collectively durations of output synthesised emphases in a target language diverge significantly from those of input natural emphases of the user. This would be quite helpful in maintaining a user's own style of emphasising words in output synthesised speech, when one builds a speaker-specific expressive speech-to-speech translator.

Analyses were not conducted at sub-word levels in this paper, because prosodic cues of a syllable or of a phone are very likely to be too language-specific. Furthermore, unlike associating words in language *A* with those in language *B* as per their meanings for cross-language prosody transfer (e.g. ***Prüfung*** and ***test***), it makes no sense to associate syllables or phones in one language with those in another (e.g. the syllables and phones of ***Prüfung*** and ***test***). It would be more appropriate to transfer prosodic cues at the word or higher level across languages and to let underlying rules of pronunciation of a target language decide how to integrate the prosodic cues into synthesis of sub-word units.

## 5. Conclusions

Analysis of duration at the sentence and word levels was conducted in this research work, based on parallel German and English speech data containing emphasis. The main motivation behind the analysis was to acquire insights into the cross-language transferability of duration of emphasised words in the context of speech-to-speech translation. According to the analysis results, the impact of emphasised words on the remaining neutral part of the same utterance was trivial. Thus it is not necessary to take advantage of duration of neutral words in input natural speech for synthesising their output translations. The analysis of recordings from different speakers suggests it is better not to transfer duration information from one speaker to another. Durations of emphasised words in input speech by themselves were not sufficient for predicting those of corresponding emphasised words in a spoken translation. However, since most of the time the hypothesis that durations of speaker-dependent emphases in the two languages obeyed the same distribution could not be rejected at $\alpha = 0.05$, durations of emphasised words in input speech would still be able to portray one's particular style of emphasising words in output synthesised speech. This finding is helpful in judging whether a personalised speech-to-speech translator reproduces a user's own way of uttering emphasis. For further research, similar analysis of pitch and intensity can be conducted and other language pairs can be involved.

## 6. Acknowledgements

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

## 7. 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] 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.

[4] 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.

[5] 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.

[6] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, and D. Jurafsky, "Modelling prominence and emphasis improves unit-selection synthesis," in *Proc. of Interspeech*, Aug. 2007, pp. 1282–1285.

[7] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "Intent transfer in speech-to-speech machine translation," in *Proc. of SLT*, Dec. 2012, pp. 153–158.

[8] 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.

[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] A. Rosenberg, E. Cooper, R. Levitan, and J. Hirschberg, "Cross-language prominence detection," in *Proc. of Speech Prosody*, Jan. 2012.

[11] D. Abercrombie, *Elements of General Phonetics*. Aldine Publishing Company, 1967.

[12] 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.

[13] W. J. Conover, *Practical Nonparametric Statistics*, 3rd ed. John Wiley and Sons, Jan. 1999.

[14] G. Marsaglia, W. W. Tsang, and J. Wang, "Evaluating Kolmogorov's distribution," *Journal of Statistical Software*, vol. 8, no. 18, pp. 1–4, Nov. 2003.