Speaking Rate Changes Affect Phone Durations Differently for Neutral and Emotional Speech

Yingming Gao
Institute of Acoustics and Speech Communication
Dresden University of Technology
Dresden, Germany
yingming.gao@mailbox.tu-dresden.de

Peter Birkholz
Institute of Acoustics and Speech Communication
Dresden University of Technology
Dresden, Germany
peter.birkholz@tu-dresden.de

***Abstract*—In this study we examined whether phone durations are affected differently when the speaking rate changes in neutral and emotional speech. To that end, we analyzed two sets of sentences: In the first set, each sentence was spoken with *explicitly* different speaking rates (slow, normal, fast) with a neutral emotion. In the second set, each sentence was spoken with seven emotions with *implicitly* different speaking rates. For each spoken sentence, the mean value and the standard deviation of phone durations were related to those of the corresponding normal or neutral counterparts. We found that the normalized relative standard deviation (NRSD) did not solely depend on the speaking rate, but also on the emotion. Based on these findings, we analyzed the listening effort and the naturalness of synthetic utterances, where the mean and the standard deviation of phone durations were modified independently of each other. For fast speech, listening effort and naturalness improved significantly, when the standard deviation of phone durations was reduced less than the mean phone durations. These results can be applied to time-compression strategies for synthetic speech, voice morphing, and realistic synthesis of emotional speech.**

***Index Terms*—speech synthesis, time compression and expansion, variability of duration, listening effort, naturalness**

I. Introduction

The timing of speech, as a prosodic feature, plays an important role for speaking styles and for the expression of vocal emotions. On the one hand, emotions affect the overall speaking rate, e.g., joyful and angry utterances are usually spoken faster than sad utterances [1] [2]. On the other hand, it is likely that also the individual timing of phones differs across emotions, even for similar speaking rates. For example, Vroomen et al. [3] showed that copying solely the time structure of an emotionally spoken sentence onto the same neutrally spoken sentence allows listeners to recognize neutral, boredom and anger. Beyond the understanding of emotional speech, the investigation of speech timing at different speaking rates has applications like speech synthesis, voice transformation, and foreign language learning.

The temporal pattern of natural fast or slow speech differs from that of normal-rate speech in various ways [4] [5]. Accordingly, a number of methods have been proposed to artificially compress or expand the duration of an utterance. Among these, proportional scaling is the simplest method, by which all segments are uniformly reduced or increased by the same degree, thus the relative temporal pattern of original speech is kept in the modified speech. However, the comprehensibility of proportionally compressed speech typically degrades, even when the rate is still below that of natural fast speech. This is not due to the speech rate per se but due to the unnatural timing [6] [7]. Covell et al. [8] proposed a method called "Math1" as an alternative to proportional scaling, which imitates the compression patterns found in natural fast speech. In this method, the components of an utterance are non-uniformly compressed across speaking rates, with pauses/silences being compressed most, and stressed vowels least. Time-compressed speech using this method was clearly preferred and better comprehended by listeners than speech modified by proportional compression. Janse et al. [4] analyzed the effect of speaking rate on *syllable* durations (as opposed to segment durations) and found that stressed syllables tend to be compressed more than unstressed syllables when speakers increase their speaking rate. However, when this finding was applied to the artificial time-compression of speech, it neither improved the intelligibility [9] [10] [11] nor the listeners subjective preference [12] over proportionally compressed speech. This is not consistent with the results achieved with the Math1 method by Covell et al [8].

Kozhevnikov and Chistovich [13] analyzed the effect of speaking rates on *word* durations and found that, regardless of speaking rates, each word accounted for a constant proportion of the duration of the whole sentence (i.e., the relative duration of words was invariable). However, they also found that the relative duration of the sounds within a word did vary as a function of speaking rate.

In essence, phone durations and syllable durations appear to scale non-proportionally when the speaking rate changes. However, with regard to the artificial change of speaking rate, there is no agreement on whether a non-proportional scaling actually improves the quality of the modified speech compared to proportional scaling. In addition, it is not clear whether increased or decreased speaking rate affects phone durations always in the same way, or whether it depends on the emotion. In the present study, we therefore analyzed how the standard deviation of phone durations changes relative to the mean phone duration for different speaking rates in both neutral and emotional utterances (Sec. II). In addition, we conducted a perception experiment to find out whether changing the

ISBN 978-90-827970-1-5 © EURASIP 2018

standard deviation of phone durations independently from the mean phone duration for faster and slower speaking rates can improve the listening effort and naturalness over proportionally compressed or expanded speech (Sec. III).

II. Analysis of Phone Durations

Here, we used two databases to analyze how phone durations change in terms of their mean and their standard deviation when the same sets of sentences are spoken with different speaking rates.

A. Neutral Speech at Slow, Normal, and Fast Speaking Rates

The first database was the Multi-Modal Annotated Synchronous Corpus of Speech (MMASCS) [14]. This corpus contains the segmented and annotated audio files of the productions of 130 German sentences, each of which was spoken at three different speaking rates: normal, fast and slow. These are 390 speech items in total. For each item, we calculated two quantities: the relative articulation rate (RAR) and the normalized relative standard deviation (NRSD) of the phone durations.

The RAR was calculated as the articulation rate of the item under consideration divided by the articulation rate of the corresponding item spoken at the normal rate for the same sentence:

$$RAR_i = \frac{N_i/t_i}{N_{normal}/t_{normal}} = \frac{\mu_{normal}}{\mu_i} \qquad (1)$$

Here, N_i, t_i and μ_i are the number of phones, the duration (excluding pauses), and the mean phone duration of the item *i* under consideration, respectively, and N_{normal}, t_{normal} and μ_{normal} are the number of phones, the duration (excluding pauses), and the mean phone duration of the corresponding normal-rate item, respectively.

The NRSD was calculated as the relative standard deviation (standard deviation divided by mean) of the phone durations of the item under consideration divided by the relative standard deviation of the corresponding normal-rate item:

$$NRSD_i = \frac{\sigma_i/\sigma_{normal}}{\mu_i/\mu_{normal}} = \frac{\sigma_i/\mu_i}{\sigma_{normal}/\mu_{normal}} \qquad (2)$$

Here, σ_i and σ_{normal} are the standard deviations of the item *i* and the corresponding normal-rate item, respectively. If phone durations would scale proportionally with a change in speaking rate, both σ_i and μ_i would change by the same factor with respect to σ_{normal} and μ_{normal}, so that $NRSD_i = 1$. Non-proportional changes of phone durations due to speaking-rate changes are accordingly reflected in NRSD values smaller or greater than 1.

Fig. 1 shows the boxplots of RAR and NRSD for the speech items. A one-way ANOVA for both RAR and NRSD with speech rate as the fixed factor revealed that the difference between the groups was statistically significant with $F(2, 387) = 6604.158$, $p < 0.001$, and $F(2, 387) = 126.913$, $p < 0.001$, respectively. In addition, multiple comparisons using paired *t*-tests (considering Bonferroni correction) suggested that the mean values of the groups significantly differed from each other for a significance level of 0.01.

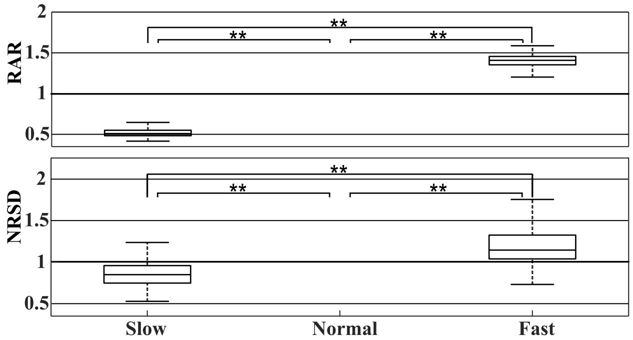

Fig. 1. Boxplots of relative articulation rate and normalized relative standard deviation of phone durations for slow, normal and fast speech. ($^{**}p < 0.01$)

As expected, the relative articulation rate is greater than 1 for the fast-speech items, and smaller than 1 for the slow-speech items. However, the NRSD values for the fast and slow items suggest that phone durations do not scale proportionally with respect to normal-rate speech. In fact, also the way of non-proportional scaling differs between fast and slow speech. For speech changed from normal to slow rate, the standard deviation of the phone durations increases less than the mean phone duration, so that the relative phone durations within one utterance tend to get more equal to each other. In contrast, for speech changed from normal to fast rate, the standard deviation of phone durations reduces less than the mean phone duration, so that differences of relative phone durations increase, i.e., compared to normal-rate speech, short phones get under-proportionally shorter, and long phones get over-proportionally longer.

B. Emotional Speech with Different Implicit Speaking Rates

To test whether the above findings are universal for faster-than-normal and slower-than-normal speech, we analyzed, as a second database, the Berlin Database of Emotional Speech (Emo-DB) [15]. This database contains the segmented and labelled utterances of 10 actors, each of whom spoke 10 sentences each in a happy, angry, anxious, fearful, bored and disgusted way as well as in a neutral version. In contrast to the MMASCS database, the different versions of the same sentences were not deliberately produced with different speaking rates, but the different speaking rates are an implicit consequence of the acted vocal emotions. In the Emo-DB, all utterances were evaluated by listeners with respect to the recognition and naturalness of the intended emotions [15]. In the present study, we used only the subset of utterances that was positively evaluated and for which also the neutral counterpart by the same speaker was positively evaluated (431 items in total).

For each of these items, the relative articulation rate and the normalized relative standard deviation of the phone durations were calculated analogously to Sec. II-A, where the normal

ISBN 978-90-827970-1-5 © EURASIP 2018

items were considered to be those spoken with the neutral emotion.

Fig. 2 shows the boxplots of RAR and NRSD for all speech items. Also here, a one-way ANOVA for both RAR and NRSD with the emotion state as the fixed factor revealed a significant difference between groups with $F(6, 424) = 65.546$, $p < 0.001$ and $F(6, 424) = 4.416$, $p < 0.001$, respectively. Furthermore, two-sample *t*-tests (due to the unbalanced sample sizes) were used to test significant differences in the mean NRSD values and the mean RAR values for all pairs of emotions, considering Bonferroni correction. For RAR, the differences were significant ($p < 0.05$) for all pairs of emotions except one. For the NRSD, the mean values significantly differed for the three pairs fear-boredom, boredom-neutral, and neutral-anger.

The results show that phone durations in faster and slower emotional speech do *not* necessarily change in the same way as in neutral speech. For example, while the NRSD is smaller than 1 for slower neutral speech (see Fig. 1), it may be greater than 1 for slower emotional speech (see "disgust" and "boredom" in Fig. 2). Furthermore, while the NRSD is greater than 1 for faster neutral speech, it may be smaller than 1 for faster emotional speech (see "fear" in Fig. 2). These results suggest that we use different strategies to adapt the phone durations for faster or slower speech depending on the emotional state. In other words, the change of the standard deviation of phone durations for faster or slower speech (with respect to normal or neutral speech) is not completely linked to the change of the mean phone duration, but it is an independent degree of freedom in the timing of speech.

III. Duration Modification Method and Perception Experiment

A. A Method for Duration Modification

From the observations and analysis in Sec. II, we can derive a simple model for the *non-proportional* adaptation of phone durations of an input utterance to an output utterance with a higher or lower speaking rate. The basic idea is to consider the durations of the phones of the input utterance and the output utterance as samples from two normal distributions and then equate the z-scores of both distributions for all phones in the utterance. The Fig. 3 shows the histogram of all phone durations of the 130 sentences at the normal speaking rate from the MMASCS database, and the best fit with the density function of a normal distribution (black curve). Even though the assumption that phone durations are normally distributed is not strictly valid (a Gamma distribution actually fits phone duration distributions slightly better [16]), we regard the fit as reasonable enough for the model to express durational characteristics.

The parameters of the distribution of the input utterance can be estimated from the sample mean μ_{in} and the sample standard deviation σ_{in} of the phone durations of the input utterance. The mean and standard deviation of the phone duration distribution of the output utterance must be given to the method as parameters μ_{out} and σ_{out}. Then a specific phone duration d_{in} of the input utterance maps to a specific phone duration d_{out} of the output utterance as follows:

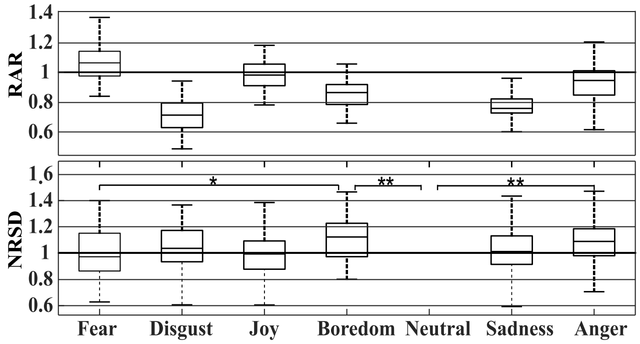

Fig. 2. Boxplots of relative articulation rate and normalized relative standard deviation of phone durations for different emotions. For the RAR, the mean values are significantly different between all pairs of emotions. (*$p < 0.05$; **$p < 0.01$)

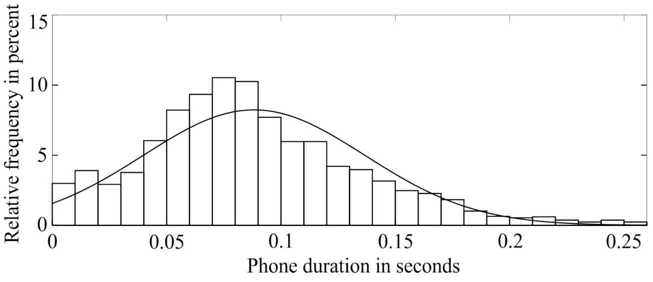

Fig. 3. Histogram of phone durations at normal rate speech and best fit with the density function of a normal distribution.

$$d_{out} = \frac{\sigma_{out}}{\sigma_{in}} \cdot (d_{in} - \mu_{in}) + \mu_{out} \tag{3}$$

This equation can also be written in terms of two factors $f_1 = \frac{\mu_{out}}{\mu_{in}}$ and $f_2 = \frac{\sigma_{out}}{\sigma_{in}}$ (i.e., $f_1 = \frac{1}{RAR}$ and $f_2 = \frac{NRSD}{RAR}$ if using normal rate speech as input), which define the relative change of the mean and the standard deviation, respectively:

$$d_{out} = f_2 \cdot (d_{in} - \mu_{in}) + f_1 \cdot \mu_{in} \tag{4}$$

Here, the special case $f_1 = f_2$ obviously corresponds to the proportional scaling of phone durations. When the standard deviation of the output phone duration increases too much, it may happen that d_{out} becomes negative. In this case, the according phone could be dropped in the output phone sequence. Here, we adopted a different alternative and set a lower duration limit of 20 milliseconds to all phones, i.e.,

$$d_{out} := max\{20ms, d_{out}\}. \tag{5}$$

B. Perception experiment

The model for duration modification proposed above was then used to test the perceptual relevance of the independent factor f_2 on the naturalness and the intelligibility of synthetic utterances with different articulation rates.

1) Creation of Stimuli: We selected five neutral basis utterances (sentences 'a01', a02', 'b01', 'b02', 'b10') of the male speaker '15' from the Emo-DB. These specific utterances were chosen because their phonetic realization by the speaker was close to the canonical form. Based on the original phone durations and fundamental frequency contours, each of the five sentences was then re-synthesized in 25 variants using the diphone speech synthesizer Mbrola [17] with the diphone database "de2" (male German speaker). The 25 variants per sentence were generated by modifying the phone durations according to (4) and (5) with all combinations of the factors $f_1 = \{0.5, \frac{1}{\sqrt{2}}, 1, \sqrt{2}, 2\}$ and $f_2 = \{0.5, \frac{1}{\sqrt{2}}, 1, \sqrt{2}, 2\}$.

Hence, each sentence was generated with five target durations and with five different standard deviations of phone durations per target duration. The range of f_1 and f_2 between 0.5 and 2 represents roughly the corresponding range found across all emotional utterances in the Emo-DB. The fundamental frequency contour was temporally stretched or compressed according to the duration factor f_1.

2) Experimental Procedure: After synthesizing all 5 x 25 = 125 stimuli, 21 native Germans (13 males and 8 females; mean age: 38.1 years) with normal hearing ability participated in a perception experiment. The listening test consisted of two sessions. In the first session, the participants were asked to evaluate the listening effort of all the stimuli on a 5-point Likert scale with "1" standing for "very high listening effort" and "5" for "very low listening effort". The stimuli were presented to the participants in random order (individual order for each participant) over closed-ear headphones (AKG K240) in a sound-proof room. The participants could repeat the presentation of each stimulus once on demand. The second session was similar to the first session, but this time the participants had to rate the naturalness of the stimuli on a 5-point Likert scale from "1" for "very unnatural" to "5" for "very natural".

C. Results

The results of the ratings of the listening effort are shown in Fig. 4. In each subplot, the horizontal axis represents the five-point scale, and the vertical axis represents the number of selections of a certain point. The vertical black lines mark the mean scores for each specific combination for f_1 and f_2. The subplots in each column have the same mean durations (i.e., speaking rates) but different standard deviations (i.e., temporal organization). A two-factor ANOVA with f_1 and f_2 as fixed factors and either subjects or sentences as repeated measures showed a significant main effect of both f_1 and f_2 on the listening effort ($F(4, 2600) = 497.5$, $p < 0.001$ and $F(4, 2600) = 2.56$, $p = 0.037 < 0.05$, respectively). The interaction effect between f_1 and f_2 was also highly significant ($F(16, 2600) = 3.37$, $p < 0.001$). In general, the listening effort reduced from fast to slow speaking rates, and the best scores were achieved for slow speech ($f_1 = 1.41$).

Furthermore, we applied paired *t*-tests to find out whether the stimuli with the same factor f_1 but with different factors f_2 were rated significantly different from the stimuli that correspond to proportional scaling (where $f_1 = f_2$). Significant differences (at a significance level of 0.05 before Bonferroni correction) are marked by stars "*" in Fig. 4. For example, the stimuli with $(f_1, f_2) = (0.5, 1.41)$ and $(f_1, f_2) = (0.5, 2)$ were rated as needing a significantly smaller listening effort than the proportionally scaled stimuli with $(f_1, f_2) = (0.5, 0.5)$. This means that for fast speech, utterances with proportionally compressed phone durations require a higher listening effort than utterances where the standard deviation is reduced less than the average phone duration.

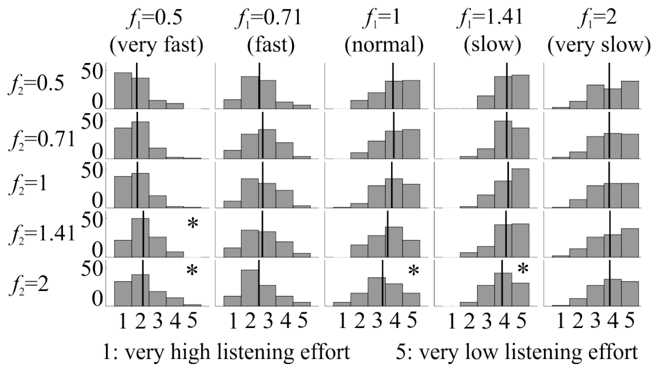

Fig. 4. Histograms of ratings for listening effort for the synthetic speech stimuli.

Similar results were obtained for the naturalness ratings, as shown in Fig. 5. Also here, the main effects of both f1 and f2 were highly significant ($F(4, 2600) = 337.46$, $p < 0.001$, and $F(4, 2600) = 4.21$, $p = 0.002 < 0.01$, respectively). The items with the normal speaking rate were found to sound most natural, and with increasing or decreasing speaking rate, the naturalness usually degraded. Separate paired *t*-tests were also carried out analogously as above to test whether non-proportional scaling improves the naturalness compared to proportional scaling for the different duration factors. Also here, the statistically significant improvements of naturalness were obtained especially for "very fast" synthetic speech when f_2 was greater than f_1.

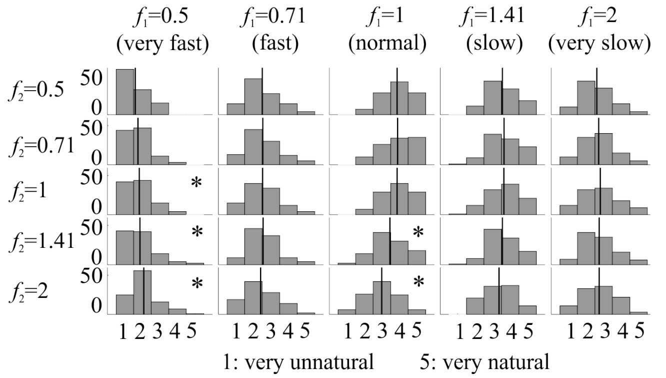

Fig. 5. Histograms for naturalness ratings for the synthetic speech stimuli.

Compared to proportional scaling, whether a non-proportional scaling actually improves the quality of the

ISBN 978-90-827970-1-5 © EURASIP 2018

modified speech can be revealed by the mean scores (vertical lines) within each column. We can always find some specific combinations of different factors f_1 and f_2 whose stimuli obtained better or comparable scores over the diagonal cases (where $f_1 = f_2$ in Fig. 4 and 5), which are analogous to the proportional scaling in [9-12]. Among these, the improvements of listening effort and naturalness were statistically significant for very fast speech.

IV. Discussion and Conclusions

The analysis of the distributions of phone durations in two databases showed that phone durations change non-proportionally when the speaking rate is changed, and that the strategy of this non-proportional change depends on the emotional state. This indicates that a change of speaking rate involves not only a change of the average phone duration, but that there is at least one additional degree of freedom that is consciously varied. In the listening experiment, we furthermore showed that fast speech generated from normal speech by proportional compression of phone durations is not optimal from the perspective of listening effort and naturalness. Instead, fast synthetic speech is more easily perceived and more natural when the standard deviation of phone durations is relatively less reduced than the mean phone duration. The proposed model for duration modification is rather simple and does not account, e.g., for individual differences of phoneme types (e.g. different compressibility of consonants classes and vowels), but it may be an appropriate starting point for speaking rate changes in synthetic speech or for voice morphing, especially for representing timing differences in emotional speech. Finally, the proposed NRSD could be a complementary means for the *analysis* of speech compression and expansion, as it considers both speaking rate and phone duration distribution.

Acknowledgment

We would like to thank the participants of the listening experiments and Felix Burkhardt and Astrid Bartels for providing the MBROLA pho-files for the utterances in the Emo-DB.

References

[1] D. Galanis, V. Darsinos, and G. Kokkinakis, “Investigating emotional speech parameters for speech synthesis,” Electronics, Circuits, and Systems, 1996. ICECS’96., Proceedings of the 3rd IEEE International Conference on. IEEE, vol. 2, pp. 1227-1230, 1996.

[2] C. Breitenstein, D. V. Lancker, and I. Daum, “The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample,” Cognition and Emotion, vol. 15, no. 1, pp. 57-79, 2001.

[3] J. Vroomen, R. Collier, and S. J. L. Mozziconacci, “Duration and intonation in emotional speech,” The 3rd European Conference on Speech Communication and Technology, 21-23 September, Berlin, Germany, Proceedings, 1993.

[4] E. Janse, A. Sennema, and A. Slis. “Fast speech timing in Dutch: durational correlates of lexical stress and pitch accent,” INTERSPEECH 2000 1st Annual Conference of the International Speech Communication Association, October 16-20, Beijing, China, Proceedings, pp. 251-254, 2000.

[5] R. F. Port, “Linguistic timing factors in combination,” The Journal of the Acoustical Society of America, vol.69, no.1, pp. 262-274, 1981.

[6] P. A. Gade, and B. M. Carol, “Listening rate and comprehension as a function of preference for and exposure to time-altered speech,” Perceptual and Motor Skills, vol. 68, no. 2, pp. 531-538, 1989.

[7] C. P. Fulford, “Can learning be more efficient?: Using compressed speech audio tapes to enhance systematically designed text,” Educational Technology, vol. 33, no. 2, pp. 51-59, 1993.

[8] M. Covell, M. Withgott, and M. Slaney. “Mach1: Nonuniform time-scale modification of speech,” Acoustics, Speech and Signal Processing, 1998, Proceedings of the 1998 IEEE International Conference on. vol. 1. IEEE, pp. 349-352, 1998.

[9] E. Janse, “Intelligibility of time-compressed speech: three ways of time-compression,” INTERSPEECH 2000 1st Annual Conference of the International Speech Communication Association, October 16-20, Beijing, China, Proceedings, pp. 786-789, 2000.

[10] E. Janse, “Comparing word-level intelligibility after linear vs. non-linear time-compression,” INTERSPEECH 2001 2nd Annual Conference of the International Speech Communication Association, September 3-7, Aalborg, Denmark, Proceedings, pp. 1407-1410, 2001.

[11] E. Janse, S. Nooteboom, and H. Quen, “Word-level intelligibility of time-compressed speech: prosodic and segmental factors,” Speech Communication, vol. 41, no. 2, pp. 287-301, 2003.

[12] E. Janse, “Word perception in natural-fast and artificially time-compressed speech,” 15th ICPhS, August, Barcelona, Spain, Proceedings, pp. 3001-3004, 2003.

[13] V. A. Kozhevnikov, and L. A. Chistovich, Speech: Articulation and perception, Washington, DC: U.S. Joint Publications Research Service, 1965.

[14] D. Schabus, M. Pucher, and P. Hoole, “The MMASCS multi-modal annotated synchronous corpus of audio, video, facial motion and tongue motion data of normal, fast and slow speech,” LREC, pp. 3411-3416, 2014.

[15] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of German emotional speech,” INTERSPEECH 2005 6th Annual Conference of the International Speech Communication Association, September 4-8, Lisbon, Portugal, Proceedings, vol. 5, pp. 1517-1520, 2005.

[16] T. H. Crystal and A. S. House “Segmental durations in connected speech signals: Preliminary results,” The Journal of the Acoustical Society of America, vol. 72.3, pp. 705-716, 1982.

[17] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Vrecken, “The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes,” Spoken Language, 1996. ICSLP 96. Fourth International Conference, Proceedings, vol. 3, pp. 1393-1396, IEEE, 1996.