Stress and Accent Transmission In HMM-Based Syllable-Context Very Low Bit Rate Speech Coding

Milos Cernak, Alexandros Lazaridis, Philip N. Garner, Petr Motlicek

Idiap Research Institute, Martigny, Switzerland

{Milos.Cernak,alaza,Phil.Garner,Petr.Motlicek}@idiap.ch

Abstract

In this paper, we propose a solution to reconstruct stress and accent contextual factors at the receiver of a very low bit-rate speech codec built on recognition/synthesis architecture. In speech synthesis, accent and stress symbols are predicted from the text, which is not available at the receiver side of the speech codec. Therefore, speech signal-based symbols, generated as syllable-level log average F0 and energy acoustic measures, quantized using a scalar quantization, are used instead of accentual and stress symbols for HMM-based speech synthesis. Results from incremental real-time speech synthesis confirmed, that a combination of F0 and energy signal-based symbols can replace their counterparts of text-based binary accent and stress symbols developed for text-to-speech systems. The estimated transmission bit-rate overhead is about 14 bits/second per acoustic measure.

Index Terms: Low bit-rate speech coding, HMM-based speech synthesis

1. Introduction

Very low bit-rate (VLBR) speech coders operate at bit-rates $\leq$ 600 bits per second (b/s). Considering the Shannon'source coding theorem, the minimal bit-rate is based on the entropy of the phonemes and basic prosodic features, together on the order of 100 b/s uncompressed; this is an "operating point" of our work. The prototype of our VLBR coder is real-time and produces intelligible speech with low communication delay[1].

For achieving this very low bit-rate, parameteric speech coding has to be used. VLBR speech coders with speech recognition and synthesis architectures belong to the most popular approaches. In our work we use phoneme recognition as a *transmitter*, and phoneme synthesis as a *receiver*. Transmission of pitch and duration of the symbols is required to recover the original prosody. While corpus-based techniques have been applied in the past for VLBR speech coding systems (e.g., [1]), HMM-based speech synthesis systems (HTS) are beneficial from an adaptation point of view, and the system footprint [2].

In our recent work on VLBR speech coding [3, 4], we expressed the aim to unify the transmitted information on the syllable context level. We hypothesised that unifying context across all levels of transmitted information may decrease the overall bit rate of the speech coder, while allowing acceptable communication delay. In this paper we focus on one module of the *receiver*, and specifically on the context reconstruction from the transmitted parameters. The context has significant impact on the naturalness of the speech generated by an integrated HTS system. We propose a solution to reconstruct stress and accent contextual factors at the receiver side. Both factors are normally predicted by a text analyzer (e.g., using a word dictionary for stress and a prediction model for accent). The text in phoneme based recognition and synthesis is not available, however, the speech signal is. Therefore our main research hypothesis is that the acoustic correlates of text-based accent and stress prediction (we call them speech signal-based) in HTS synthesis can perform as well as text-based contextual factors. For validating the hypothesis, we use the quantized average log pitch (F0) and energy labels in the syllable context HTS modelling.

The idea of using speech signal-based labels has recently been applied to low bit rate F0 coding [5]. More closely to our work, F0 quantized symbols were used for unsupervised labelling of accentual context in [6]. The F0 sequence for synthesis from text was generated using an average voice model. We extended the idea of quantized F0 by using multiple signal-based symbols (F0 and energy), and passing from phone-level to syllable-level quantization that fits our very low bit rate speech coding framework. The novel aspect of our work is a combination of different signal-based measures that aim to replace the conventional text-based contextual factors that are not available at the receiver side of the codec, such as factors related to word and phrase contextual factors. While the effectiveness of quantized F0 symbols was proved in [6] for a syllable-timed language, we found that using a combination of more acoustic measures is necessary for stress-timed languages (see description in the next section). We evaluated the accent and stress modelling for English language.

The paper is structured in the following way. The next Sec. 2 introduces the stress and accent contextual factors and their acoustic signal-based correlates. The Sec. 3 presents an analysis of predictability power of acoustic correlates based on the normalised mutual information measure. Sec. 4 describes the experiments and results, and finally Sec. 5 concludes the paper with the discussion and future work outline.

2. Stress and accent contextual factors

In this work, by the term *stress*, we refer to the lexical stress of a word, which is the stress placed on syllables within words. In some languages, such as English, a word might have a secondary stress, distinct from the primary one. Languages can be classified into two main categories, syllable-timed and stress-timed languages, depending on whether the duration of every syllable is equal or the duration of the intervals between two stressed syllables are equal, respectively [7, 8]. In early work, Fry [9, 10] showed that stress was highly correlated to acoustic variations in speech, such as increased duration and higher intensity and moreover higher pitch values. Gordon [11], showed that in stress-timed languages, such as English, stresses assigned at the word level are correlated to speech properties such

[1]Some examples available at https://www.idiap.ch/project/recod/demo/incremental-coding/

as an increased duration of the stressed phonemes or syllables, higher intensity values, and/or hyper-articulation, rather than with pitch variations on these syllables. Additionally, Sluijter and van Heuven [12] concluded that, in English, pitch is correlated to phrase-level pitch accent rather than to word-level stress. The controversy between these two above-mentioned views is caused by the fact that Fry's research was conducted on words positioned on parts of phrases that conveyed not only word-level stress but also phrase-level prominence, leading to the conclusion that stress was also correlated to higher pitch values.

On the other hand, *accent* refers to the phrase- or sentence-level prominence given to a syllable. The syllables conveying phrasal prominence are often called pitch accented syllables. In the rest of the paper, with the term accent we refer to the pitch accent, which is the phrasal prominence distinct from tones correlated to the boundaries of intonational phrases. Pitch accents mainly convey semantic information such as focus and emphasis. In some cases, a stressed syllable can be promoted to pitch accented syllable based just on it's position in a phrase or on the focus/emphasis the speaker intends to give to the specific part of the sentence to convey a specific message [11].

In our work, we used syllable-based F0 and energy acoustic measures as acoustic correlates of the accent and stress features. Although the duration was also reported as an important measure, we omitted duration from this work. The duration of syllable is correlated to the number of phonemes in a syllable, a contextual factor already used in the HTS training.

The accent and stress HTS contextual features could take binary values that represent accented and/or stressed syllable. We generated F0 p_i and energy e_i labels by averaging their values per syllable, and quantizing the average values using a scalar quantization (see Sec. 4.2). The quantization code-book, created per speaker, was linearly spaced between the $\mu-3\sigma$ and $\mu+3\sigma$ boundaries, where μ is the mean and σ is the standard deviation of all measurements belonging to the training data of a particular speaker. So the quantization book was created per acoustic measure and speaker. To simplify the creation of the question set used for context clustering of these signal-based labels, 3-bit quantization was used resulting in 8 different labels for each acoustic measure.

3. Analysis of information related to signal-based labels

This section describes the analysis of signal-based labels, based on Mutual Information (MI) of the text-based and signal-based labels. The work in [6] showed that quantized F0 are effective for accent prediction for a syllable-timed language. The main goal of this section is to estimate the predictability power of individual and combined quantized F0 and energy contextual factors for English, a stress-timed language.

Conventional *stress* and *accent* contextual factors are predicted from the text to be synthesised, and the binary labels $c \in 0, 1$ are assigned to the current and previous syllable in a phrase. The proposed average quantisized F0 p_i and energy e_i features are calculated from the speech signal measures, and the labels $M_i \in \{0, 1, \ldots, 7\}$ were assigned also to the current and previous syllable (8 labels resulted from use of 3-bit code-books). To combine the p_i and e_i features (to capture, e.g., higher pitch and lower energy) into a single label, we constrained $M_i = p_i = e_i$, where the value of the label in question was the same for both acoustic measures. It simplified the construction of the question set for the context clustering as well.

The MI as a measure of the conventional *stress* and *accent* labels C and speech signal measure-based M_i labels can be defined as:

$$I(C; M_i) = \sum_{c} \sum_{m \in M_i} p(c, m) \log_2 \left(\frac{p(c, m)}{p(c)p(m)} \right), \quad (1)$$

where $p(c, m)$ is a joint probability of C and M_i, and $p(c)$ and $p(m)$ are their marginal probabilities. The MI is a measure of information in bits that conveys M_i about C, and it is normalised by the mutual information measure with the entropy of C defined as:

$$H(C) = -\sum_{c} p(c) \log_2 p(c). \quad (2)$$

We evaluated the normalised measure $\frac{I(C;M_i)}{H(C)}$ for the following classes (options) of C:

1. C_a, where $C_a =$ *accent*, and is a measure of information in bits that conveys M_i about conventional *accent* labels,
2. C_s, where $C_s =$ *stress*, and is a measure of information in bits that conveys M_i about conventional *stress* labels,
3. $C_{s \wedge a}$, where $C_{s \wedge a} =$ *stress* $\wedge$ *accent*, and is a measure of information in bits that conveys M_i about the intersection of conventional *stress* and *accent* labels, i.e., $C_a \cap C_s = \{c : c \in C_a \wedge c \in C_s\}$.
4. $C_{s|a}$, where $C_{s|a} =$ *stress*|*accent*, and is a measure of information in bits that conveys M_i about the union of conventional *stress* and *accent* labels, i.e., $C_a \cup C_s = \{c : c \in C_a \text{ or } c \in C_s\}$,

Table 1 shows the normalised MI values of the male `bdl` testing speaker from the CMU-ARCTIC speech database [13].

Table 1: *Normalised MI values of* `bdl` *voice.*

$\frac{I(C;M_i)}{H(C)}$	`bdl`		
	p_i	e_i	p_i, e_i
$C = C_a$	10.1	8.2	16.7
$C = C_s$	10.9	7.4	18.3
$C = C_{a \wedge s}$	11.4	9.6	17.9
$C = C_{a\|s}$	12.4	8.5	**20.6**

From the normalised MI values we see that (a) p_i is more informative than e_i and (b) individual p_i and e_i values are less predictive than their combination. The best predicted syllables using the signal-based labels are those which are either accented or stressed, i.e., those represented by conventional *stress* or *accent* label class $C_{a|s}$. From this analysis we can conclude that proposed M_i labels have enough predictability power to replace the conventional text-based stress and accent labels. In that way we do not model explicitly stress and accent. Rather we hypothesise, that the context clustering does the job in a data-driven manner, and selected contextual questions about M_i values will reflect the information about actual conventional stress and accent features.

4. Experimental setup

The experimental setup we used is depicted in Fig. 1. Focusing on evaluation of the contextual factors of the HTS modelling that is at the receiver side, we abstracted the transmitter

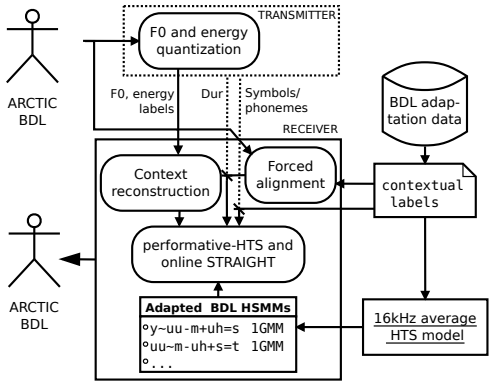

Figure 1: *VLBR speech coding experimental setup with recognition-synthesis architecture, abstracting the transmitter (dotted lines) except for F0 and energy quantization.*

side and used the true input to the receiver, i.e., phoneme sequence from the manually prepared syllable context labels, duration of phonemes from the forced alignment against natural speech, and original pitch values.

As our objective was to build an incremental real-time speech coder, we used the incremental p-HTS system speech parameter generation with 2 previous label smoothing [14]. Speech samples were finally re-synthesized using a real-time incremental STRAIGHT re-synthesis [15]. In this way speech samples are generated frame by frame with each processed phoneme.

4.1. Contextual factors

Our recent work on contextual factors in HTS concluded that the context above syllable (the word and phrase context) is less important [3]. To evaluate the proposed syllable-context signal-based labels, we have designed 3 kinds of synthesis systems integrated within the receiver:

1. **Baseline**: No accent/stress context, the system used only the full phonetic context,
2. **Conventional**: Accent/stress context, the system extended the baseline system with the conventional accent and stress labels,
3. **Proposed**: signal-based label context , the system extended the baseline system with the proposed signal-based labels.

Tab. 2 lists contextual factors for all the three systems.

The combination of signal-based qunatized F0 p_i and quantized energy e_i factors for context clustering was achieved by listing all possible values of p_i and e_i within one question, with constrain $M_i = p_i = e_i$ (as described in Sec. 3).

4.2. Generation of signal-based labels

For training and testing of the baseline and conventional systems, the contextual labels provided by the CMU-ARCTIC database were used. For the proposed system, new labels were generated as follows.

First, the code-books for average log F0 and log energy were created. All log F0 measurements of the training set were extracted using the TEMPO method of [16] using 5 ms frame shift. The syllable boundaries were extracted from the contextual labels provided by the CMU-ARCTIC database, and average values were calculated per syllable. Fig. 2 shows the histogram of calculated values together with the boundaries for the 3-bit scalar quantization $(\mu - 3\sigma, \mu + 3\sigma)$ for the `bdl` speaker. Similarly, the syllable-average log energy values were calculated using the signal-processing tracter tool [17] and the log energy code-book was created.

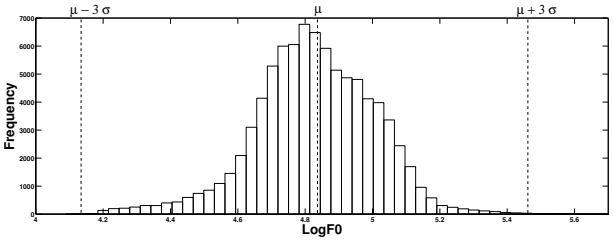

Figure 2: *A histogram of the syllable-average log F0 values for the code-book creation.*

In the next step, all F0 and energy measurements were quantized and new signal-based labels were created for both training and testing sets. We replaced the conventional stress and accent labels with these new labels (see Tab. 2), and used them for training and testing of the proposed system.

4.3. HTS training

For building the HMM models, the HTS v. 2.1 toolkit [18] was used. Specifically, the implementation from the EMIME project [19] was taken. The speech data which were used had 16kHz sampling frequency. Five-state, left-to-right, no-skip HSMMs were used. The speech parameters which were used for training the HSMMs were 39 order cepstral coefficients, log-F0 and 21-band aperiodicities, along with their delta and delta-delta features, extracted every 5 ms.

For the experiments, the CMU-ARCTIC database was used. For each of the three systems presented in this work (baseline, conventional and proposed) an American English average model was built using five speakers, including 3 males (`bdl`, `jmk` and `rms`) and 2 females (`clb` and `slt`). The unilex [20] phone-set was used, consisting of 41 phones. Each utterance with unique id was assign to:

- the training set, if $0 \leq id \leq 450$,
- the test set, if $450 < id \leq 500$, and
- the adaptation set, if $id > 500$.

In this way, the training set of the average model contained 4493 sentences (some corrupted utterances were excluded from the training). The `bdl` speaker was selected as a testing speaker. The `bdl` adaptation set of 131 sentences was used for adapting each of the three average models, resulting into three models used for the three different systems described above. Finally the `bdl` test set of 100 sentences was used for evaluating the three systems.

4.4. Evaluation and results

For the subjective evaluation of the three systems, two ABX tests were conducted. According to [21], the ABX test is suitable for rating small degradation using a continuous impairment scale, and expert (trained) listeners should be used. In our case we were interested if there is some small degradation of the proposed system, comparing to the conventional system.

Table 2: *Description of contextual factors used in the proposed syllable-context synthesis systems. The abbreviations PS and CS stand for the previous and current syllable, respectively.*

Baseline	**Conventional**	**Proposed**
– full phoneme context – the number of phonemes in the PS – the number of phonemes in the CS – vowel name in the CS	– full phoneme context – the number of phonemes in the PS – the number of phonemes in the CS – vowel name in the CS – whether the PS stressed $\{0, 1\}$ – whether the PS accented $\{0, 1\}$ – whether the CS stressed $\{0, 1\}$ – whether the CS accented $\{0, 1\}$	– full phoneme context – the number of phonemes in the PS – the number of phonemes in the CS – vowel name in the CS – quantized F0 p_i label of the PS $\{0, \ldots, 7\}$ – quantized energy e_i label of the PS $\{0, \ldots, 7\}$ – quantized F0 p_i label of the CS $\{0, \ldots, 7\}$ – quantized energy e_i label of the CS $\{0, \ldots, 7\}$

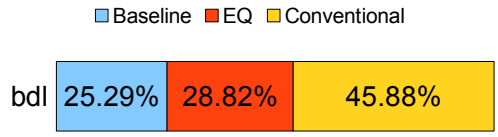

Figure 3: *ABX subjective evaluation test results (in percentages) for the comparison between the baseline and conventional systems, for* `bdl` *speaker.*

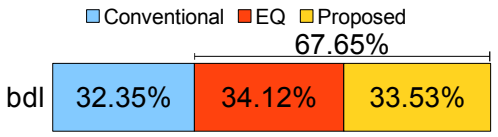

Figure 4: *ABX subjective evaluation test results (in percentages) for the comparison between the conventional and proposed systems, for* `bdl` *speaker.*

In the first ABX test, listeners were asked to choose between speech samples coming from two systems, the baseline and the conventional systems. The motivation behind the first test was to investigate whether the use of stress and accent information based on the textual labels will improve the quality of the synthesized speech. In the second ABX test, listeners were asked to choose between speech samples produced from the conventional and the proposed systems. The motivation for the second ABX test was to see whether the use of stress and accent information based on the speech signal on the encoder's part rather than based on textual labels, will effect the overall quality of the synthetic speech on the decoder's side.

In both tests 17 listeners, most of them from Idiap speech group, participated. In each test, the listeners were asked to listen for each pair of sentences the two samples (as many times as they wanted), and choose between the two samples in terms of the overall quality. Additionally, the listeners could choose a third option, "both samples sound the same", if they had no preference between them. For both tests the same 10 sentences from the test set were used [2], concluding to 30 samples in total, i.e., 10 samples for each of the three systems, baseline, conventional and proposal respectively.

Fig. 3 presents the results of the first test. The conventional system significantly outperforms the baseline system. These results clearly show that, as it was expected, the use of stress and accent information based on the textual labels, i.e., the conventional labels, improve the quality of the synthesized speech.

Fig. 4 presents the results of the second listening test. The results obtained indicate that the performances are not significantly different, i.e., the proposed system with labels inferred from speech performed similarly as the conventional system developed to text-to-speech synthesis with labels inferred from text. As stated above, the motivation for the second ABX test was to investigate if the use of stress and accent information based on the speech signal on the transmitter's part will deteriorate the overall quality of the synthetic speech on the decoder's side in respect to the system using stress and accent as conventional text-based contextual factors. Under this hypothesis, the comparison we are focused on in the second ABX test is between the preference score of the conventional system and the combination of the scores of the EQ case and the proposed system (i.e., we classify EQ case as success). From this point of view, the proposed system outperforms the conventional one in terms of preference score, with 67.65% over 32.35%, respectively. These results clearly validate our hypothesis that speech signal-based stress and accent features could perform as well as text-based contextual factors.

The `bdl` test set contained 1312 syllables in 274 seconds of speech. In average, there were so 4.8 of syllables per second. Using the 3-bit quantization book, each acoustic measure needed around 14 b/s that we may consider as the transmission overhead per acoustic measure.

5. Conclusions

The results clearly validate our hypothesis that speech signal-based stress and accent features could perform as well as text-based contextual factors. Moreover, listeners more prefered the proposed system with quantized syllable-level log average F0 and energy acoustic measures, than conventional stress and accent features as provided by the TTS front-end. The proposed solution works well for speech coding and is probably less effective for speech synthesis from text.

We believe that adding new syllable-based acoustic measures, such as voice quality measures, that might correlate with some other prosodic or emotional features can further improve the naturalness of the encoded speech.

6. Acknowledgements

This research was supported by the RECOD project by armasuisse, the Procurement and Technology Center of the Swiss Federal Department of Defence, Civil Protection and Sport and under the SIWIS project by the Swiss National Science Foundation.

[2] A subset of voice samples is available at `http://www.idiap.ch/project/recod/demo/stress-encoding`.

7. References

[1] G. V. Baudoin and F. El Chami, "Corpus based very low bit rate speech coding," in *Proc. of ICASSP*, vol. 1. IEEE, Apr. 2003, pp. I–792–I–795 vol.1. [Online]. Available: http://dx.doi.org/10.1109/icassp.2003.1198900

[2] K. Tokuda, T. Masuko, J. Hiroi, T. Kobayashi, and T. Kitamura, "A very low bit rate speech coder using HMM-based speech recognition/synthesis techniques," in *Proc. of ICASSP*, vol. 2. IEEE, May 1998, pp. 609–612 vol.2. [Online]. Available: http://dx.doi.org/10.1109/icassp.1998.675338

[3] M. Cernak, P. Motlicek, and P. N. Garner, "On the (UN)importance of the contextual factors in HMM-based speech synthesis and coding," in *Proc. of ICASSP*. IEEE, May 2013, pp. 8140–8143. [Online]. Available: http://dx.doi.org/10.1109/icassp.2013.6639251

[4] M. Cernak, X. Na, and P. N. Garner, "Syllable-Based Pitch Encoding for Low Bit Rate Speech Coding with Recognition/Synthesis Architecture," in *Proc. of Interspeech*, Aug. 2013, pp. 3449–3452. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2013/i13_3449.html

[5] T. Nose and T. Kobayashi, "Very low bit-rate F0 coding for phonetic vocoder using MSD-HMM with quantized F0 context," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, May 2011, pp. 5236–5239. [Online]. Available: http://dx.doi.org/10.1109/icassp.2011.5947538

[6] T. Nose, K. Ooki, and T. Kobayashi, "HMM-based speech synthesis with unsupervised labeling of accentual context based on F0 quantization and average voice model," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, Mar. 2010, pp. 4622–4625. [Online]. Available: http://dx.doi.org/10.1109/icassp.2010.5495548

[7] K. L. Pike, *The intonation of American English*. University of Michigan Publications in Linguistics 1. Ann Arbor: University of Michigan Press, 1945.

[8] D. Abercrombie, *Elements of general phonetics*. Edinburgh: Edinburgh University Press, 1967.

[9] D. Fry, "Duration and intensity as physical correlates of linguistic stress," *Journal of the Acoustical Society of America*, vol. 27, pp. 765–768, 1955.

[10] ——, "Experiments in the perception of stress," *Language and Speech*, vol. 1, pp. 120–152, 1958.

[11] G. Matt, *Disentangling stress and pitch accent: Toward a typology of prominence at different prosodic levels*. Oxford University Press: in Harry van der Hulst (ed.). To appear, In Word Stress: Theoretical and Typological Issues, 2014.

[12] M. S. Agaath and J. v. H. Vincent, "Spectral balance as an acoustic correlate of linguistic stress," *Journal of the Acoustical Society of America*, vol. 100, pp. 2471–2485, 1996.

[13] J. Kominek and A. Black, "The CMU Arctic speech databases," in *Proc. of 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223 – 224.

[14] M. Astrinaki, N. d'Alessandro, B. Picart, T. Drugman, and T. Dutoit, "Reactive and continuous control of HMM-based speech synthesis," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, Dec. 2012, pp. 252–257. [Online]. Available: http://dx.doi.org/10.1109/slt.2012.6424231

[15] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *MAVEBA*, 2001.

[16] H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proc. of Eurospeech*, Budapest, Hungary, 1999.

[17] P. N. Garner and J. Dines, "Tracter: a lightweight dataflow framework," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 1894–1897.

[18] "HMM-based speech synthesis system version 2.1," 2010. [Online]. Available: http://hts.sp.nitech.ac.jp

[19] M. Wester, J. Dines, M. Gibson, H. Liang, Y.-J. Wu, L. Saheer, S. King, K. Oura, P. N. Garner, W. Byrne, Y. Guan, T. Hirsimäki, R. Karhila, M. Kurimo, M. Shannon, S. Shiota, J. Tian, K. Tokuda, and J. Yamagishi, "Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project," in *SSW*, 2010, pp. 192–197. [Online]. Available: http://www.isca-speech.org/archive/ssw7/ssw7_192.html

[20] S. Fitt, "Documentation and user guide to unisyn lexicon and post-lexical rules," Center for Speech Technology Research, University of Edinburgh, Tech. Rep., Tech. Rep., 2000.

[21] V. Grancharov and W. . B. Kleijn, "Speech Quality Assessment," in *Springer Handbook of Speech Processing*, J. Benesty, Sondhi, and Y. Huang, Eds. Springer Berlin Heidelberg, 2008, pp. 83–100. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-49127-9_5