**University of
Zurich**^{UZH}

**Zurich Open Repository and
Archive**

University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2015

Stable and Unstable Intervals as a Basic Segmentation Procedure of the Speech Signal

Glavitsch, Ulrike ; He, Lei ; Dellwo, Volker

Abstract: The concept of acoustically stable and unstable intervals to structure continuous speech is introduced. We present a method to compute stable intervals efficiently and reliably as a bottom-up approach at an early processing stage. We argue that such intervals stand in close relation to the rhythm of speech as they contribute to the overall temporal organization of the speech production process and the acoustic signal (stable intervals = intervals of reduced movement of certain articulators; unstable intervals = intervals of enhanced movement of certain articulators). To test the relationship of stability intervals with speech rhythm we investigated the between-speaker variability of stable and unstable intervals in the TEVOID corpus. Results revealed that significant between-speaker variability exists. We hypothesize from our findings that the basic segmentation of speech into stable and unstable intervals is a process that might play a role in human perception and processing of speech.

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: <https://doi.org/10.5167/uzh-128615>
Conference or Workshop Item

Originally published at:

Glavitsch, Ulrike; He, Lei; Dellwo, Volker (2015). Stable and Unstable Intervals as a Basic Segmentation Procedure of the Speech Signal. In: INTERSPEECH 2015, Dresden, 6 September 2015 - 10 September 2015. s.n., 31-35.

Stable and Unstable Intervals as a Basic Segmentation Procedure of the Speech Signal

Ulrike Glavitsch¹, Lei He², Volker Dellwo²

¹ EMPA, Swiss Federal Laboratories for Materials Science and Technology, Switzerland

² Department of Comparative Linguistics, University of Zurich, Switzerland

ulrike.glavitsch@empa.ch, lei.he@uzh.ch, volker.dellwo@uzh.ch

Abstract

The concept of acoustically stable and unstable intervals to structure continuous speech is introduced. We present a method to compute stable intervals efficiently and reliably as a bottom-up approach at an early processing stage. We argue that such intervals stand in close relation to the rhythm of speech as they contribute to the overall temporal organization of the speech production process and the acoustic signal (stable intervals = intervals of reduced movement of certain articulators; unstable intervals = intervals of enhanced movement of certain articulators). To test the relationship of stability intervals with speech rhythm we investigated the between-speaker variability of stable and unstable intervals in the TEVOID corpus. Results revealed that significant between-speaker variability exists. We hypothesize from our findings that the basic segmentation of speech into stable and unstable intervals is a process that might play a role in human perception and processing of speech.

Index Terms: stable intervals, speech rhythm, speaker variability, speech processing

1. Introduction

There are different ways to structure the speech signal. Most common are segmentations in units of phonemes or syllables performed by tools such as Praat [1]. Automatic alignment algorithms recognize phoneme boundaries using an underlying ASR system provided a phonetic transcription of the spoken text is given [2, 3, 4]. Methods to detect syllable boundaries automatically are reported in [5, 6, 7]. The segmentations are used in different areas of speech research and processing. Phonetic alignments are widely used in phonetic analysis [8]. In addition, they deliver the training data for the acoustic models of ASR systems [9]. Furthermore, durational and intensity measures are computed from phonetic alignments to recognize speakers [10].

A more elementary way to segment the speech signal is by focusing on predominantly bottom-up characteristics of the signal. This is particularly relevant for structuring signals when no linguistic knowledge is available. Such methods might thus teach fundamental ways of processing complex acoustic signals like speech. One such method, for example, is to divide speech into a sequence of filled and unfilled (pause) intervals as it is implemented in the software SpeechIndexer [11, 12]. An automatic built-in pause finder detects the speech pauses [13]. SpeechIndexer's pause finder is applicable for all languages, in fact, it has been used for documenting and transcribing endangered Formosan aboriginal languages [11].

Furthermore, voiced and unvoiced parts of speech are often distinguished and there exist a number of voiced/unvoiced decision algorithms in different contexts [14, 15].

In this paper, we present stable and unstable intervals as another way to segment speech and we test whether this type of segmentation shows systematic variability between speakers. Stable intervals are defined as voiced sections with a quasi-equal energy or a quasi-flat envelope. Unstable intervals, however, may show significant energy rises or decays. Stable and unstable intervals (in general: stability intervals) have been presented as anchor points of detection in the context of a fundamental frequency (F_0) estimation algorithm [16].

Why did we test the variability of stability intervals between speakers? Speech production is a complex process underpinned by sophisticated neuromuscular activities for the control of speech organs [17]. The movements of speech organs, like the movements of other parts of the body (see [18, 19] for leg movements in human gait), are highly idiosyncratic. This is because individuals vary in the physical characteristics of their articulators and thus develop different strategies to operate their movements. It has been demonstrated that it is very likely that such movement idiosyncrasy directly influences the prosodic temporal organization of speech, i.e. the speaker's individual rhythm [20, 21]. Previous studies revealed, for example, that temporal characteristics of vocalic, intervocalic, voiced, or peak-to-peak intervals showed significant between-speaker effects even when within-speaker variability is strong [10, 20, 21]. Evidence has also been delivered that this is true for proportional between-syllable intensity characteristics which are also likely to be highly influenced by speaker specific movements of the articulators [22]. It seems likely that intervals over which speech is stable or unstable are also to some degree influenced by speaker specific movement behavior. It seems plausible that acoustic stability might be a result of reduced movement activity of the articulators responsible for keeping the amplitude envelope constant while acoustic variability should correlate with enhanced movement activity of these articulators. For this reason, we hypothesize that the proportional time over which speech is stable (%Stable) reveal between-speaker variability in a comparable way as other temporal prosodic characteristics that are influenced by movement behavior of the articulators. Further, it is possible that articulators requiring larger movement activity to change from one stable region to the next might be characterized by a higher proportion over which speech is unstable (%Unstable). The variability of stable and unstable interval durations in terms of their standard deviation (Δ Stable and Δ Unstable) is also expected to be influenced by

idiosyncratic movements. Speakers revealing more complex movement activity over consonant clusters might thus have a higher variability in the durations of unstable intervals.

In section 2 and 3 we describe the computation and evaluation of stability intervals in speech (stable, unstable, voiceless). In section 4 we investigate between-speaker variability of the measures and in section 5 we discuss possible roles of stable and unstable intervals for human speech perception and processing. We draw conclusions and give an outlook to future work in section 6.

2. Computation of stable intervals

We propose different ways to compute stable and unstable intervals automatically: (1) a simple method, (2) an ideal approach and (3) a method that attempts to approximate the ideal approach. We show that method (3) is better in terms of error rates than the simple method but far more efficient than the ideal system.

As mentioned in the introduction, stable intervals are voiced regions in the speech signal with quasi-constant energy. In order to compute stable intervals automatically, we first calculate the energies at regular time steps in the speech signal. For this purpose, the speech signal is divided into frames of 20 ms length that overlap by 10 ms. There are several ways to compute the energy of a frame. One possibility is to compute the energy on a fixed-size short-time window. This is surely the easiest and most straight-forward method. Another approach is to compute it on a pitch-synchronous window in case of a voiced frame and on a fixed-size window in case of an unvoiced frame. In this second case, the energies are divided by the length of the window in order to compare them due to the varying-sized window lengths. Thus, the compared energies are in fact mean energies or signal powers (energy per time). The second approach is the ideal way to compute stable intervals, as the mean energies are always computed on the precise F_0 period. For example, the mean energy of a speech signal with a pitch of 75 Hz computed on pitch-synchronous window of 13.333 ms is more meaningful than computed on a fixed-sized window of 20 ms. In the latter case, 1.5 periods fit into the fixed-sized window which is certainly not optimal. The problem with pitch-synchronous energy computation is that F_0 is not available at an early processing stage.

Thus, we propose a third way to compute the energies at regular time intervals that attempts to approximate the pitch-synchronous energy calculation. In this third method, the mean energies are computed on a scale of different window lengths each of which corresponds to a different pitch period length or a multiple of it. Then, an optimization step is performed to find the best window length. This method is similar to pitch-scaled harmonic filtering (PSHF) where an optimal window length is calculated for finding harmonic and non-harmonic spectra [23]. The scale of window lengths corresponds to fundamental frequencies of 50, 55, 60, ..., 95 Hz. For example, exactly 4 periods of a signal with pitch period corresponding to 220 Hz fit in a window of 55 Hz, 4.5 periods of a signal with pitch period of 230 Hz would fit in a window of 55 Hz or 3.5 periods of it would fit into the next smaller window of 60 Hz. Both window sizes are not optimal in this latter case but the portion of the signal not fitting exactly an integer number of times in the window is small compared with the window length. After computing the mean energy for the different window sizes the optimal window size is selected as

the one where the mean energies of a number of successive frames (two frames before the center frame, the center frame and two frames after the center frame) show the least variation.

Given the mean energies for each frame we classify a frame as voiced or unvoiced and for a voiced frame as either stable or unstable as follows. A frame is voiced if its mean energy exceeds a given threshold, the absolute height of the frame's maximum or minimum peak is above some level and the number of zero crossings per millisecond is less than a given number. A voiced frame is defined as *stable* if the mean energies between the preceding and the current frame and between the current and the following frame vary by less than 50 %. A stable interval is therefore a sequence of stable frames, an unstable interval a sequence of unstable frames. Fig. 1 shows the speech signal containing the words "the north". The voiced frames of the signal are shown at the bottom. There are three stable intervals S_0 , S_1 and S_2 of lengths 1, 9 and 27 in this voiced region.

Figure 1. Three stable intervals S_0 , S_1 and S_2 in the voiced part of a speech signal.

3. Evaluation of methods

The two methods where stable intervals are computed based on pitch-scaled energy computation (PITCH-SCALED) and based on a fixed-size window length (FIXED-SIZE) are evaluated against the ideal method (baseline system) where stable intervals are calculated using pitch-synchronous window lengths. The baseline system uses a fundamental frequency estimation algorithm that has shown to achieve very low gross pitch error (GPE) rates on a clean speech database [16]. The FIXED_SIZE method uses the whole frame length of 20 ms as window length.

For comparison of the three methods, the voiced/unvoiced classification is the same for all three approaches to compute stable and unstable intervals. The three methods are evaluated on two different databases. The first database is the Keele pitch reference database for clean speech [24]. 5 male and 5 female speakers read Aesop's fable "The north wind and the sun" in English. The second database is the TEVOID (Temporal Voice Idiosyncrasy) corpus that contains read and spontaneously spoken sentences from 16 native speakers of Zürich German [10]. There are 8 female and 8 male speakers that read 256 sentences and speak 16 sentences spontaneously.

We measured the stable errors and the unstable errors of both methods PITCH-SCALED and FIXED-SIZED with respect to the baseline system. A stable error occurs if a frame is classified as unstable where a stable frame exists. On the other hand, an unstable error is counted if a frame is classified as stable instead as unstable. In other words, a stable error corresponds to a miss of a stable frame whereas an unstable error is a false alarm. The stable error rate is the percentage of stable errors with respect to the total number of stable frames of the baseline method. The unstable error rate is defined analogously.

Table 1 shows the results of the two methods PITCH-SCALED and FIXED-SIZED with respect to the baseline method for the Keele pitch reference database. We list the stable error rates (SE), the unstable error rates (UE) and the total error rate (TE) for all the speakers as well as for all male and all female speakers separately. The total error rate is the sum of stable error rate and unstable error rate. Table 2 shows the corresponding error rates computed for the three groups of speakers for the TEVOID corpus [10] with respect to the ideal baseline system.

	SE (%)	UE (%)	TE (%)
All speakers			
PITCH-SCALED	8.61	16.27	24.88
FIXED-SIZE	6.07	21.15	27.22
Female speakers			
PITCH-SCALED	5.15	13.21	18.36
FIXED-SIZE	4.38	18.57	22.95
Male speakers			
PITCH-SCALED	11.89	19.25	31.14
FIXED-SIZE	7.68	23.67	31.35

Table 1. Keele database results: Error rates of two stable intervals computation methods.

	SE (%)	UE (%)	TE (%)
All speakers			
PITCH-SCALED	7.37	16.22	23.59
FIXED-SIZE	5.20	21.76	26.96
Female speakers			
PITCH-SCALED	5.56	15.97	21.53
FIXED-SIZE	4.13	22.75	26.88
Male speakers			
PITCH-SCALED	9.25	16.49	25.74
FIXED-SIZE	6.33	20.68	27.01

Table 2. TEVOID corpus results: Error rates of two different stable intervals computation methods.

The results show that the PITCH-SCALED method outperforms the FIXED-SIZE method on both databases for all groups of speakers in terms of total errors (TE). The difference is up to 5.35 % in case of the female speaker group on the TEVOID corpus. The lowest difference of only some tenths of percent is achieved for the male speaker group on the Keele database.

The FIXED-SIZE method has a lower stable error rate but its unstable rate is significantly higher for all speaker groups. The absolute differences of the unstable error rate (UE) between the two methods vary between 4.19 % (male speakers on TEVOID) and 6.78 % (female speakers on TEVOID).

The method PITCH-SCALE is not optimal what the stable errors are concerned especially for the male voices where we observe stable error rates (SE) of 11.89 % for the Keele database and 9.25 % for TEVOID. However, we consider a worst-case recognition rate of almost 90 % and a worst-case false alarm rate of less than 20 % sufficient for our purposes. Besides, the PITCH-SCALE method is very efficient compared with the ideal baseline system that requires F_0 estimates.

4. Between-speaker variability of stability

To address the question of between-speaker variability we processed temporal characteristics of stable and unstable intervals as introduced in section 1 (%Stable, %Unstable, Δ Stable and Δ Unstable) between speakers in the TEVOID corpus [10]. The variability measures, Δ Stable and Δ Unstable, were log transformed to normalize for overall rate influences on the variability that might vary between speakers [20].

Figure 2 shows boxplots of %Stable for read sentences and spontaneous speech for the same 16 speakers of the TEVOID corpus. The figure reveals that there is considerable variability between speakers in both speaking styles but that, unlike other temporal measures based on phonemic segments [20, 21], the between-speaker variability seems to be of a different quality in spontaneous speech compared to read sentences.

Figure 2. Box plots of %Stable for read (top) and spontaneous sentences (bottom) of TEVOID.

Factorial ANOVAs were run on all of these measures (R code: `aov(measure~speaker*speakingStyle, data=data)`). For both %Stable and %Unstable, significant interactions were found (%Stable: $F_{\text{Speaker}[15,4314]} = 38.934$, $p < 0.0001$, $F_{\text{Style}[1,4314]} = 56.092$, $p < 0.0001$, $F_{\text{Speaker*Style}[15,4314]} = 8.329$, $p < 0.0001$; %Unstable: $F_{\text{Speaker}[15,4314]} = 33.01$, $p < 0.0001$, $F_{\text{Style}[1,4314]} = 34.88$, $p < 0.0001$, $F_{\text{Speaker*Style}[15,4314]} = 6.31$, $p < 0.0001$) which supports the impression from Figure 2. The interactions make the interpretation of the main effects difficult.

Given that the between-speaker variability is different in the spontaneous speaking style it seems inevitably the case the

varying lexical content with its phonological/phonotactic characteristics contributed highly to this variability. For this reason we looked at the between sentence variability of %Stable for the first 15 sentences of the TEVOID corpus in the read speech. Figure 3 reveals that this variability can be very large and consistent between sentences. A one-way ANOVA confirmed this picture by showing a highly significant main effect ($F[14, 225]=41.9, p<0.001$).

Figure 3. Boxplot of %Stable for the first 15 read sentences in the TEVOID corpus produced by 16 speakers each.

For both $\Delta\text{StableLn}$ and $\Delta\text{UnstableLn}$, the speaker effect was found to be significant, but the speaking style was not (%Unstable: $F_{\text{Speaker}}[15,4314] = 27.462, p < 0.0001$, $F_{\text{Style}}[1,4314] = 0.088, p > 0.7$; $\Delta\text{UnstableLn}$: $F_{\text{Speaker}}[15,4314] = 33.64, p < 0.0001$, $F_{\text{Style}}[1,4314] = 0.078, p > 0.7$). We found significant interactions again which make the main effects hard to interpret. Significant speaker effects, however, seem likely ($\Delta\text{StableLn}$: $F_{\text{Speaker*Style}}[15,4314] = 2.396, p < 0.002$; $\Delta\text{UnstableLn}$: $F_{\text{Speaker*Style}}[15,4314] = 4.307, p < 0.0001$).

5. Role of stable intervals

As shown in section 4 stable intervals vary between speakers and between sentences of a language which is why they might be of relevance for the processing of speaker specific and linguistic information in the speech communication process.

Furthermore, we propose stable intervals as starting or anchor points in automatic speech processing. The speech signal in stable intervals is easier and more reliably recognized than in unstable intervals. In particular, the speech signal in high-energy stable intervals shows mostly clear and regular signal periods. (A signal period is the length of that portion of the signal that is grossly similar to its neighbor section in voiced speech. In high-energy stable intervals not only the periods but also the forms of the signal between neighboring periods are often similar.) Thus, automatic speech processing best starts from such stable intervals and explores the other areas in a second step using the prior knowledge from the stable intervals. This procedure would be applicable for all kinds of speech processing tasks, e.g. pitch detection, formant estimation, automatic speech alignment and automatic speech recognition in general. From stable intervals where the recognition task is easy and reliably to perform the detection would continue to unstable intervals. The recognition in unstable intervals, however, might profit from the knowledge

in the stable intervals. The search space can be drastically reduced if values in neighboring stable intervals are already known.

A concrete speech processing task, namely a pitch-detection algorithm, has been implemented using this strategy [16]. It finds reliable pitch estimates on the high-energy stable intervals first and propagates them to unstable intervals in a second step. Pitch estimates on stable intervals are computed with a straight-forward method, i.e. by comparing distances between maximum peaks and their heights. Propagation goes to both the left and right side of a stable interval and from high-energy to low-energy regions. A look-ahead of two periods is used to compensate for the irregularities of periods in unstable regions. The algorithm is easy to understand and very efficient as it operates in the time-domain only. Furthermore, it achieves very low gross pitch error rates and is comparable to other state-of-the art pitch detection methods.

The encouraging results of the pitch detection algorithm manifest the assumption that recognition processes in human speech perception run in a similar way. Stable and unstable intervals may also serve as a model for low-level human speech perception. Not all portions of the speech signal are equally well perceived by humans. It is possible that the portions of the signal that are easy to detect are recognized predominantly over regions during which detection is more complex.

6. Conclusions and outlook

We presented acoustically stable and unstable intervals as a new way to segment the speech signal. We described a method to compute such intervals reliably and efficiently in a bottom-up manner at an early stage of processing. It achieves nearly 90 % coverage of stable intervals with respect to an ideal method and a false alarm rate of less than 20 %.

Furthermore, we showed that the percentage of stable intervals (% Stable) and the mean difference of stable interval length ($\Delta\text{StableLn}$) vary significantly between speakers and between sentences. This means that stability regions are influenced by the linguistic content of a sentence and possibly contribute to the individuality of voices.

Finally, stable and unstable intervals have shown to be ideal starting or anchor points for speech processing algorithms as they serve as regions where recognition tasks can easily and reliably be performed. For the same reason, stable and unstable intervals may also play a role in human speech perception.

It would be interesting to investigate stability intervals in more detail in the future for a variety of phonetic research tasks, in particular for speaker recognition. Given that suprasegmental characteristics of speech contribute to speaker identity, the temporal characteristics derived from stability intervals may possibly improve speaker recognition.

7. Acknowledgements

This research was funded by Grant Number GRS-027/13 of the Gebert Rf Stiftung and Grant 135287 of the Swiss National Science Foundation. We wish to thank Adrian Leemann and Marie-Jos Kolly for their work in the creation of the TEVOID database.

8. References

- [1] P. Boersma, D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 5.4.08, retrieved 24 March 2015 from <http://www.praat.org/>.
- [2] J.-Ph. Goldman, "EasyAlign: an automatic phonetic alignment tool under Praat," in *INTERSPEECH 2011 – 12th Annual Conference of the International Speech Communication Association, August 28 – 31, Firenze, Italy, Proceedings*, 2011, pp. 3233–3236.
- [3] S. Brognaux, S. Roekhaut, T. Drugman, R. Beaufort, "Train&Align: A new online tool for automatic phonetic alignment," *Spoken Language Technology Workshop (SLT), 2012 IEEE*, Miami, FL, USA, pp. 416–421.
- [4] S. Hoffmann, B. Pfister, "Text-to-Speech Alignment of Long Recordings Using Universal Phone Models," in *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association, August 25 – 29, Lyon, France, Proceedings*, 2015, pp. 1520–1524.
- [5] H. R. Pfister, S. Burger, S. Heid, "Syllable detection in read and spontaneous speech," in *Proc. ICSLP 1996*, vol. 2, pp. 1261–1264.
- [6] Z. Xie and P. Niyogi, "Robust acoustic-based syllable detection," in *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, September 17-21, Pittsburgh, PA, USA, Proceedings*, 2006, pp. 1571–1574.
- [7] J. Sirigos, N. Fakotakis, G. Kokkinakis, "A hybrid syllable recognition system based on vowel spotting," *Speech Communication*, vol. 38, pp. 427–440, 2002.
- [8] J. Harrington, "Phonetic Analysis of Speech Corpora," Wiley-Blackwell, 2010.
- [9] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, February, 1989.
- [10] V. Dellwo, A. Leemann, M.-J. Kolly, "Speaker idiosyncratic rhythmic features in the speech signal," in *INTERSPEECH 2012 – 13th Annual Conference of the International Speech Communication Association, 9 – 13 September, Portland, Oregon, USA, Proceedings*, 2012.
- [11] J. Szakos, U. Glavitsch, "Seamless Speech Indexing and Retrieval: Developing a New Technology for the Documentation and Teaching of Endangered Formosan Aboriginal Languages," in *EISTA'14 – International Conference on Education and Information Systems: Technology and Applications*, Orlando, USA, July 21 – 25, pp. 88–93, 2004.
- [12] U. Glavitsch, K. Simon, J. Szakos, "SpeechIndexer – A Flexible Software for Audio-Visual Language Learning," in *icEIC 2011 - Intl. Conference on Education, Informatics and Cybernetics*, Orlando, USA, pp. 79–82, 2011.
- [13] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," *The Bell System Technical Journal*, vol. 54, pp. 297–315, 1975.
- [14] E. Fisher, J. Tabrikian, S. Dubnov, "Generalized likelihood ratio test for voiced/unvoiced decision using the harmonic plus noise model," *Proc. ICASSP*, pp. 440 - 443, 2003.
- [15] R. G. Bachu, S. Kopparthi, B. Adapa, B. D. Barkana, "Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy," K. Elleithy (ed.), *Advanced Techniques in Computing Sciences and Software Engineering*, Springer Netherlands, pp. 279–282, 2010.
- [16] U. Glavitsch, K. Simon, "A Cognition-Oriented Approach to Fundamental Frequency Estimation," *NLPCS 2013 - 10th International Workshop on Natural Language Processing and Cognitive Science*, Marseille, France, October 15 – 16, pp. 105–116, 2013.
- [17] V. Dellwo, M. Huckvale, M. Ashby, "How is individuality expressed in voice? An introduction to speech production and description for speaker classification," In: C. Müller (ed), *Speaker Classification I*, Berlin and Heidelberg: Springer Verlag, 2007, pp. 1–20.
- [18] D. S. Matovski, M. S. Nixon, S. Mahmoodi, J. M. Carter, "The effect of time on the performance of gait biometrics," In: *Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)*, Washington D.C., 2010, pp. 1–6.
- [19] J.-H. Yoo, D. Hwang, K.-Y. Moon, M. S. Nixon, "Automated human recognition by gait using neural network," In *First Workshops on Image Processing Theory, Tools and Applications*, Sousse, 2008, pp. 1–6.
- [20] V. Dellwo, A. Leemann, M.-J. Kolly, "Rhythmic variability between speakers: articulatory, prosodic, and linguistic factors," *J. Acoust. Soc. Am.* 137, pp. 1513–1528, 2015.
- [21] A. Leemann, M.-J. Kolly, V. Dellwo, "Speech-individuality in suprasegmental temporal features: implications for forensic voice comparison," *Forensic Sci. Int.* 238, pp. 59–67, 2014.
- [22] L. He, V. Dellwo, "Speaker idiosyncratic variability and intensity across syllables," in *Proceedings of INTERSPEECH 2014*, Singapore, pp. 233–237.
- [23] S. Roa, M. Bennewitz, S. Behnke, "Fundamental frequency estimation based on pitch-scaled harmonic filtering," *Proc. of ICASSP*, pp. 397–400, 2007.
- [24] F. Plante, G. F. Meyer, W. A. Ainsworth, "A pitch extraction reference database," *Proc. Eurospeech '95*, pp. 837–840, 1995.