Syllabification for Afrikaans speech synthesis

Daniel R. van Niekerk
Multilingual Speech Technologies, North-West University,
Vanderbijlpark, South Africa.
Email: daniel.vanniekerk@nwu.ac.za

***Abstract*—This paper describes the continuing development of a pronunciation resource for speech synthesis of Afrikaans by augmenting an existing pronunciation dictionary to include syllable boundaries and stress. Furthermore, different approaches for grapheme to phoneme conversion and syllabification derived from the dictionary are evaluated. Cross-validation experiments suggest that joint sequence models are effective at directly modelling pronunciations including syllable boundaries. Finally, some informal observations and demonstrations are presented regarding the integration of this work into a typical text-to-speech system.**

***Index Terms*—pronunciation dictionary, syllabification, Afrikaans, speech synthesis, text-to-speech**

I. Introduction

Recent work on text-to-speech (TTS) synthesis has applied increasingly sophisticated statistical and machine learning techniques to extract and generate patterns from appropriate speech corpora, i.e. corpus-based acoustic modelling and speech generation [1], [2]. In principle, some of these techniques are flexible enough to perform acoustic modelling directly (to various degrees) from the natural language text, thereby potentially bypassing intermediate signal or linguistic representations [3], [4]. Some of the advantages of these approaches include avoiding uncertainties or errors in intermediate representations, the possibility of optimising larger parts of systems directly, and reducing the cost associated with the development of intermediate resources (such as pronunciation dictionaries). In practice, however, acoustic modelling for TTS usually involves several components and representations of both the audio signal (such as spectral envelope, fundamental frequency and segment durations) and text (typically words, syllables and phones). The above-mentioned machine learning techniques are then used to model and generate the signal components from the linguistically motivated features [5]. While complicating the task of system-wide performance optimisation, intermediate representations can make acoustic modelling more tractable and allow for more explicit control over these aspects.

This paper describes ongoing work towards augmenting lexical pronunciation resources for TTS in Afrikaans. One part of such resources is the lexicon, which for the purpose of TTS applications should ideally contain the following information [6]:

1) Orthographies and their pronunciations.
2) Part of speech (POS) and relevant semantic information to be used for disambiguation of entries.

Relevant lexical pronunciations (1) may include phone sequences (segmental), syllable boundaries and syllable stress and/or tone (suprasegmental). In addition to the lexicon, it is necessary to have an accurate method for predicting out-of-vocabulary (OOV) words, since general purpose TTS systems are not usually expected to be restricted to a closed set of words at run-time.

The specific contributions of this work are:

1) A description of work on adapting the pre-existing Resources for Closely Related Languages Afrikaans Pronunciation Dictionary (RCRL APD) [7] for Afrikaans TTS, including the definition of a syllabification protocol and addition of manually verified syllable boundaries.
2) An evaluation of models derived from the dictionary to predict new pronunciations (phones and syllable boundaries).
3) Informal observations and a demonstration of initial work of adding syllable stress information and integrating the updated resources into an Afrikaans TTS system.

The above points are described in the next three sections respectively, followed by Section V where future work is proposed.

II. Pronunciation dictionary development

As mentioned in the point (1) above, this work is based on the pre-existing RCRL Afrikaans pronunciation dictionary. This dictionary was developed by extending the vocabulary of the relatively small Lwazi pronunciation dictionary [8] with some of the most frequently occurring words in the 60 million word *Taalkomissie Korpus* compiled by the Afrikaans Language Commission [7].

The adaptation of the RCRL dictionary for application in TTS systems involved two distinct processes. Firstly, the phone set and entries of the source dictionary were modified by merging or splitting some existing phones. This was done to clarify syllable distinctions and to improve the consistency of pronunciations in similar contexts (e.g. compound word constituents and related simplex entries). In some cases reduced pronunciations were restored to more “phonemic” alternatives. Secondly, syllable boundaries were included by applying a rule-based syllabification algorithm on decompounded versions of all entries and manually correcting the output. These two processes and brief analyses of the results are presented in more detail in the following subsections.

978-1-5090-3334-5/16/$31.00 ©2016 IEEE

A. Phone specification and conventions

Towards adapting the source dictionary for application in TTS, the following overarching principles were followed:

1) Phone representations should tend towards being phonemic, assuming that acoustic models (rather than the dictionary) are the most appropriate place to capture phonetic details from the corpus.
2) More specific pronunciation forms should be preferred over reduced forms, assuming it is easier to adapt the resource to different requirements by deriving reduced forms than vice versa.
3) Phone representations that clarify the syllable structure such as diphthongs and affricates are preferred over possible alternative forms.

TABLE I
LISTS OF MAPPED PHONES (INTRA-SYLLABLE)

Merged vowels (diphthongs)	Merged consonants (affricates)	Schwa epenthesis	
/ɑː͡i/	/t͡ʃ/	/rəm/	coda
/ɔː͡i/	/d͡ʒ/	/rən/	coda
/ɔ͡i/		/ləm/	coda
/a͡i/		/fən/	coda
/i͡u/		/pəs/	onset in loanwords
/u͡i/		/dəl/	onset in loanwords
		/məb/	onset in loanwords
		/rəw/	onset in loanwords

Applying the above principles, the following specific processes and transformations were applied to the source dictionary.[1]

1) Vowel transitions are not modelled explicitly (this is inherited from RCRL), e.g.:
   - *viool* → /fi.uəl/ ↛ /fi.juəl/
   - *italiaans* → /i.ta.li.ɑːns/ ↛ /i.ta.li.ɦɑːns/
2) For entries containing diacritics affecting syllable boundaries (diaeresis and circumflex) or non-digraphic vowel sequences, the syllable boundary is generally retained and the first of adjacent vowels is usually short, e.g.:
   - *aangeë* → /ɑːn.xi.ə/ ↛ /ɑːn.xiə.ə/,
   - *afgeleë* → /af.xə.li.ə/ ↛ /af.xə.liə/,
   - *brûe* → /brœ.ə/ ↛ /brœː/, and
   - *haelstorm* → /ɦɑː.əl.stɔ.rəm/ ↛ /ɦɑːl.stɔrm/.
3) Repeated consonants are retained around compound boundaries, except in the case of common (high-frequency) words[2] and prepositional compounds, e.g.:
   - *aankoopprys* → /ɑːn.kuəp.prəis/, but
   - *middag* → /mə.dax/ ↛ /məd.dax/, and
   - *agterruit* → /ax.tə.rœyt/ ↛ /ax.tər.rœyt/.

TABLE II
DICTIONARY G2P WORD ERROR RATES (%)

	D&R	JSM-6
RCRL	9.79	7.22
TTS	9.39	6.27

4) Sequences of vowels that occur in the same syllable are merged to form a single diphthong (see Table I).
5) Sequences of consonants belonging to the same syllable are merged to form a single affricate (see Table I).
6) A schwa is inserted in some consonant clusters according to typical pronunciations, thereby forming an additional syllable (see Table I).
7) Dictionary entries were analysed for consistency, e.g. simplex entries and compound constituents were compared to ensure similar pronunciations. Changes here mostly involved restoring reduced vowels (schwa) to more specific vowels.

The analysis of entry consistency (7) involved estimating grapheme-to-phoneme (G2P) models using joint-sequence-models (JSM) [9] and checking the dictionary against predictions. Table II shows the average 10-fold cross-validation word error rates (WER) over 5 runs for the two dictionaries using JSM (N-gram length 6) and Default&Refine (D&R) models [10]. For the JSM model $1 \leq N \leq 10$ was scanned, with $N = 6$ the point where the WER asymptotes.

B. Syllable specification and conventions

A syllable may be defined broadly as a unit of organisation for a sequence of speech sounds and is often analysed as a sub-unit of morphemes or words, leading to the task of *hyphenation*, or syllabification on the "orthographic level". Syllable units may also be analysed based on their role in speech production: describing which sounds are grouped together to form units of production. While there is a relationship between the morphological and phonotactic analyses of syllables, syllable boundaries derived from these different premises may not correspond exactly for a particular word. For example hyphenation of the English word *learning* will result in *learn-ing* while considering the phonotactics and maximal onset principle the syllables should be /lɜːɹ.nɪŋ/. The specification of syllables on the "phonemic level" in the pronunciation dictionary will naturally follow the latter conventions.

The two main principles guiding the placement of syllable boundaries for Afrikaans followed here are [11], [12]:

1) A **single vowel** (or diphthong) is allowed per syllable and all syllables must contain such (i.e. there are no syllabic consonants in Afrikaans; the obligatory syllable nucleus consists of a vowel or diphthong).
2) **Maximal onset**: Subject to phonotactic and morphological constraints described below, consonants between vowels should be assigned to the following syllable.

Point (1) constrains the problem to one of finding the location of the syllable boundary between consecutive vowels,

[1] The focus here is on the differences compared to the source dictionary, for complete definitions such as phone sets refer to the resources available at: https://github.com/NWU-MuST/afr_ttsdict

[2] This is not currently based on any specific corpus, but could in principle be.

978-1-5090-3334-5/16/$31.00 ©2016 IEEE

TABLE III
ONSET CLUSTERS FOUND IN THE DICTIONARY

Allowable onset cluster	Exceptions found	Comments
Typically not split (exceptions to "maximal onset")		
/bl/	–	Probably also: compounds, sub-
/br/	compounds, sub-	
/dr/	–	None expected (due to "final-devoicing")
/dw/	–	None expected (due to "final-devoicing")
/fl/	compounds, af-, half-, hoof-, self-, -lik, -loos	
/fr/	compounds, af-, half-, hoof-, self-	
/kl/	compounds, -lik, -loos	
/kr/	compounds	
/kw/	–	Probably also: compounds
/pl/	compounds, op-, -lik, -loos	
/pr/	compounds, op-	
/tr/	compounds, ont-, uit-	
/tw/	–	Probably Also: compounds
/vr/	–	None expected (due to "final devoicing")
/xl/	compounds, -lik	Probably also: -loos
/xr/	compounds	
/sk/	compounds, des-, dus-, eens-, eers-, mis-, trans-	
/sp/	compounds, mis-	
/sw/	–	Probably also: compounds
/st/	compounds, mis-	
/skr/	compounds, mis-	
/spl/	compounds, mis-	
/str/	compounds	Probably also: mis-
Typically split (exceptions where "not split")		
/kn/	Examples: -knip, -knie, -knoop, ...	
/sf/	Examples: -sfeer	Rare: Only in sfeer, -sferies
/sl/	Examples: -slaan, -slag, -sluit, ...	
/sm/	Examples: -smaak, -smeek, -smeer, ...	
/sn/	Examples: -snaar, -snel, -snit, ...	
/spr/	Examples: -spraak, -sprei, -spring, ...	
Single consonant onsets (exceptions to "maximal onset")		
C	compounds, agter-, alles-, an-, anders-, daar-, her-, hier-, hiper-, hoof-, in-, on-, onder-, op-, van-, ver-, vol-, voor-, wan-, waar-, -af, -agtig, -in, -of, -om, -onder, -op	Often in prepositional compounds

as a result the trivial case of adjacent vowels always takes a syllable boundary. The application of point (2) resolves the boundary placement given a consonant cluster between vowels, but requires two forms of additional information [11]:

1) A list of allowable syllable onset clusters, and,
2) The locations of morpheme boundaries that affect syllable structure (and thus potentially also pronunciation).

Relevant morpheme boundaries (2) occur in compound words between compounds (sometimes after the interfixes *-s-* or *-e-*) and at prefixes and suffixes. All cases of compound boundaries are considered to constrain the placement of syllable boundaries (i.e. can override the maximal onset principle). Affixes, however, can be divided into inflectional and derivational types, where only derivational affixes typically constrain syllable boundaries.

For the purpose of disambiguating possible syllable boundary placements, allowable onset clusters may be divided into two broad categories: (1) clusters that only occur stem-initially and (2) clusters that occur more freely in morphemes or words (an example, amongst others, of the latter are clusters that form as part of inflectional affixes). For the first category, when applying the maximal onset principle, these clusters are **usually split**, except when forming an onset of a specific (known) stem. For the second category, these clusters are **usually not split**, except when a relevant morpheme boundary intervenes. Table III list the onset clusters occurring in the dictionary in these two categories and lists exceptional cases and comments.

Given the description above and the information about consonant clusters and morpheme boundaries in Table III, entries in the dictionary were manually corrected from initial rule-based annotations. In the following section syllabification and pronunciation prediction strategies are evaluated on the dictionary.

III. PRONUNCIATION PREDICTION

A. Syllabification models

As motivated in Section I, it is essential that the lexical pronunciation resource be extended to unseen words. Given the description of the syllable conventions in the previous section, it is expected that a successful syllabification model will need

TABLE IV
SYLLABIFICATION WERs

Model	WER (%)
RUL	9.42
CC-1	2.90
JSMSYL-1	86.71
JSMSYL-2	19.99
JSMSYL-3	3.83
JSMSYL-4	2.68
JSMSYL-5	2.35
JSMSYL-6	2.28

TABLE V
G2P+SYLLABIFICATION WERs (%)

N	JSM-N → CC-1	JSM-N → JSMSYL-N	JSMG2PS-N	DECOMP→ JSMG2PS-N
1	89.29	97.11	97.58	97.58
2	41.93	51.70	69.62	69.18
3	16.65	17.26	25.50	25.10
4	10.90	10.24	12.96	12.89
5	9.31	8.47	9.30	9.50
6	**8.98**	**8.03**	7.77	8.09
7	8.90	7.97	7.21	7.57
8	8.89	7.96	**7.00**	**7.40**
9	8.88	7.95	6.96	7.35

to capture the general phonotactic constraints (allowable onset clusters), a finite set of affixes that affect syllable boundaries and be able to recognise compound boundaries.

In this section the expected performance of a few syllabification models are evaluated by determining the 10-fold cross-validation WER on the newly adapted dictionary. The experiment assumes that the phone sequence is known (from the dictionary) and only inserts syllable boundaries. The following models and algorithms were tested:

- **Phonotactic rule-based** (RUL): This is a set of rules using allowable onset clusters and the sonority scale to constrain and split consonant clusters between consecutive vowels. The algorithm was adapted for Afrikaans from a description of syllabification for English [13] and was implemented in a previous TTS project [14].
- **Cluster splitting classifiers** (CC): This is set of cluster splitting classifiers (one for each cluster length) where the local phone context is used to capture syllable boundary decisions. As a minimum, the consonant cluster phones and adjacent vowels are used as features (including vowel/consonant labels), with the option of extending the feature context further left and right. Different classifiers and context sizes can be experimented with, however, for this work the context size was limited to one phone beyond the encapsulating vowels and a random forest classifier was used [15].
- **JSM syllable boundary model** (JSMSYL): JSM models were trained to represent a mapping between input phones and two symbols representing the existence or absence of a syllable boundary.

The results (mean WER over 5 runs) of the cross-validation experiment can be seen in Table IV. The rule-based result may be interpreted as a "baseline" result estimating the error rate when no information about relevant morpheme and compound boundaries are available. The CC model is essentially a local model with the chosen context size being just large enough to capture most of the exceptions as a result of morphemes listed in Table III excepting longer compound constituents. The JSMSYL results settle at lower error rates for larger N presumably because of reoccurring (longer) compound constituents. However, it is difficult to conclude that the JSMSYL model will perform better in general than the CC model, since unlike the set of morphemes in Table III which is theoretically closed (and relatively small), this is not the case for compound constituents. The dictionary also likely contains "domain-dependent" sets of compound words which may result in optimistic cross-validation results.

B. Pronunciation prediction models

A more relevant criterion is the expected error rate for the complete pronunciation prediction process (in this case G2P including syllabification – G2P+S). In this section a few simple strategies are evaluated, cascaded G2P and syllabification models are compared to a direct JSM G2P+S model (JSMG2PS). Finally, the effect of a simple (conservative) decompounder combined with the JSM G2P+S model is measured.

For the direct JSM models, the output sequences simply contained syllable boundary symbols together with phones. The decompounder implementation takes a vocabulary as input, recursively collects sub-string sequences using a procedure similar to longest string matching and chooses the sub-string sequence using a simple heuristically defined score function rewarding sub-strings in the known vocabulary. This procedure allows for finding constituents connected with OOV interfixes and a simple post-processing step is performed to merge such OOV segments with in-vocabulary segments to suit the syllabification application.

As done in previous sections, the mean 10-fold cross-validation WERs are determined for different models and strategies. Table V presents results. Columns 1 and 2 are cascaded G2P and syllabification models reported on in the previous sections (D&R → CC-1 not shown in the table resulted in a comparable WER of 11.7 %). Column 3 contains the results for the direct JSM model, with column 4 the cascaded decompounder and direct JSM model. Rows increase with N used in the associated JSM models as indicated in the column headings (bold entries indicate where WER starts to asymptote). As expected, the direct model results in the lowest measured cross-validation error. In the current experimental setup the decompounder vocabulary was sourced from the training set only, nevertheless, a positive effect can be seen in the lower N range and a low error rate overhead at higher N (compared to the direct JSM model).

IV. TTS integration

The ultimate measure of success would be an improvement in the perceived quality of and/or enhanced control over speech synthesis output. This is difficult to measure since it depends to some extent on the desired application and is constrained by properties or limitations of the speech corpus used for acoustic modelling. Speaker, accent and register variation as well as corpus size may play a role both in synthesis output and corresponding perception of quality.

In this section an informal comparison is presented between two systems: (1) built using the original RCRL dictionary with minimal adaptation (affricate and diphthong mappings as listed in Table I) paired with the rule-based syllabification algorithm (`RUL`) compared with (2) the updated dictionary and syllabification models presented here with experimental (unverified) inclusion of syllable stress information.[3]

For the TTS systems the *HTS toolkit* version 2.3beta and associated demonstration scripts[4] [16], [17], [18] in combination with the *HTS engine* version 1.09[5], modified to perform mixed excitation synthesis [19] was used to train HMM-based statistical parametric acoustic models from the Afrikaans TTS corpus developed during the *Lwazi2* project [14]. The acoustic model contexts and tree tying questions are similar to those used in the HTS demonstration scripts (as in [5]), with the only exceptions being that no "accented" word or ToBI features are used and in the case of the "gpos" feature a simplified label set marks only content and function word distinctions.

A number of sentences were taken from online sources (mostly newspapers) during the first week of September 2016 and synthesised using both systems. An informal comparison yields a few (subjective) observations:

- In some sentence pairs the rhythm or flow of the utterance sounds more natural when using the original RCRL dictionary with rule-based syllabification and without explicit stress labels.
- Individual words are sometimes clearer in the stress marked system.
- In some cases syllabification and pronunciation errors are noticeable only in the RCRL rule-based system, e.g.:
  - *verower* → /fə.ruə.vər/ instead of /fər.uə.vər/
  - *konsternasie* → /kɔn.ˈstær.nɑː.si/ instead of /kɔn.stər.ˈnɑː.si/

This informal comparison demonstrates the difficulty of evaluating components such as the pronunciation dictionary by means of perceived relative quality of a small number of utterances based on a single speech corpus. The seemingly better flow of sentences based on the RCRL rule-based system may suggest that the certain decisions such as syllable boundaries in prepositional compounds should be revisited (to conform to rather than override the maximal onset principle) for a more informal register, however, it is not yet certain whether a single decision in this regard would be suitable for both formal and informal registers. In the following section a suggestion on how to proceed when applying the current resource given corpus-specific variations such as this is briefly discussed.

Word-centric acoustic models (removing phrase and utterance contextual features) were also trained for the two systems above to investigate the synthesis of individual words. Specifically the effect of certain dictionary convention decisions (especially points 2 and 3 in Section II-A) and the degree to which the included stress information allows for perceptible manipulation of word stress patterns. Some positive results were obtained, however, further improvement of the dictionary may be possible. All generated speech samples (utterances and words) are available online at: https://github.com/NWU-MuST/afr_ttsdict

V. Conclusion and future work

The continuing development of a pronunciation resource for speech synthesis of Afrikaans by augmenting an existing pronunciation dictionary to include syllable boundaries and stress has been described. A detailed syllabification protocol was presented and applied and methods for extending this resource to unseen words were evaluated. The resulting dictionary and pronunciation prediction components were integrated into a TTS system to demonstrate the effect on speech output. The rendering of individual words are improved in some cases with additional control over the specification of syllable stress.

Future work should involve a more rigorous inclusion of stress information in the dictionary, including considering the interaction of phone specification with syllable stress (e.g. reduced forms). Once this has been done pronunciation prediction should be re-evaluated and further work on the effective integration of compound boundaries in the process may be justified.

Further development may also attempt to document and implement systematic transformations that would make it more appropriate under different conditions involving register (style of delivery) or speaker accent variation. An example is the specific question raised about applying the maximal onset principle in prepositional compounds in the previous section. This would allow a TTS developer to selectively apply sets of transformations to broadly reduce the mismatch between the pronunciation resource and a particular corpus, possibly using objective measures such as mel-cepstral distance [20] to focus on specific aspects of speech quality. This possibility is the main motivation for point (2) of the development principles outlined in Section II-A.

VI. Acknowledgement

This work was partially funded by the Department of Arts and Culture of the Government of South Africa.

[3] Obtained by adapting the output of the rule-based pronunciation prediction algorithm implemented in http://espeak.sourceforge.net/. The result was that each syllable was associated with one of three distinct labels representing the two stress levels (*primary* and *secondary*) or *un-stressed*.

[4] http://hts.sp.nitech.ac.jp/

[5] http://hts-engine.sourceforge.net/

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] Z. H. Ling, S. Y. Kang, H. Zen, A. Senior, M. Schuster, X. J. Qian, H. M. Meng, and L. Deng, "Deep Learning for Acoustic Modeling in Parametric Speech Generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, May 2015.

[3] K. Tokuda and H. Zen, "Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 4215–4219.

[4] W. D. Basson and M. H. Davel, "Comparing grapheme-based and phoneme-based speech recognition for Afrikaans," in *Proceedings of the 23rd Annual Symposium of the Pattern Recognition Association of South Africa*, Pretoria, South Africa, 2012, pp. 144–148.

[5] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proceedings of the IEEE Workshop on Speech Synthesis*, 2002, pp. 227–230.

[6] S. Fitt and K. Richmond, "Redundancy and productivity in the speech technology lexicon - can we do better?" in *Proc. Interspeech*, Pittsburgh, Pennsylvania, USA, Sept. 2006, pp. 165–168.

[7] M. H. Davel and F. De Wet, "Verifying pronunciation dictionaries using conflict analysis," in *Proc. Interspeech*, Makuhari, Japan, Sept. 2010, pp. 1898–1901.

[8] M. H. Davel and O. Martirosian, "Pronunciation dictionary development in resource-scarce environments," in *Proc. Interspeech*, Brighton, UK, Sep. 2009, pp. 2851–2854.

[9] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, May 2008.

[10] M. H. Davel and E. Barnard, "Pronunciation prediction with Default&Refine," *Computer Speech & Language*, vol. 22, no. 4, pp. 374–393, Oct. 2008.

[11] W. Daelemans, "GRAFON: A grapheme-to-phoneme conversion system for Dutch," in *Proceedings of the 12th Conference on Computational Linguistics (COLING)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1988, pp. 133–138.

[12] G. Bouma and B. Hermans, "Syllabification of Middle Dutch," in *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH)*, Lisbon, Portugal, Nov. 2012, pp. 27–38.

[13] T. A. Hall, "English syllabification as the interaction of markedness constraints," *Studia Linguistica*, vol. 60, no. 1, pp. 1–33, 2006.

[14] K. Calteaux, F. de Wet, C. Moors, D. R. van Niekerk, B. McAlister, A. Sharma Grover, T. Reid, M. Davel, E. Barnard, and C. van Heerden, "Lwazi II Final Report: Increasing the impact of speech technologies in South Africa," Council for Scientific and Industrial Research, Pretoria, South Africa, Tech. Rep. 12045, February 2013.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python ," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modelling for speech recognition," *Journal of the Acoustic Society of Japan*, vol. 21, no. 2, pp. 79–86, 2000.

[17] T. Toda and K. Tokuda, "A Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[18] H. Zen, K. Oura, T. Nose, J. Yamagishi, S. Sako, T. Toda, T. Masuko, A. W. Black, and K. Tokuda, "Recent development of the HMM-based speech synthesis system (HTS)," in *Proceedings of th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Sapporo, Japan, 2009, pp. 121–130.

[19] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proceedings of EUROSPEECH*, Aalborg, Denmark, 2001, pp. 2263–2266.

[20] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer Voice Quality of New Languages Calibrated with Mean Mel Cepstral Distortion," in *The First International Workshop on Spoken Language Technologies for Under-resoured languages*, Hanoi, Vietnam, 2008.