INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Part-of-Speech Tagging and Chunking in Text-to-Speech Synthesis for South African Languages

Georg I. Schlünz, Nkosikhona Dlamini, Rynhardt P. Kruger

Human Language Technology Research Group
CSIR Meraka Institute, Pretoria, South Africa
gschlunz@csir.co.za, ndlamini3@csir.co.za, rkruger@csir.co.za

Abstract

Text-to-speech synthesis can be an empowering communication tool in the hands of the print-disabled or augmentative and alternative communication user. In an effort to improve the naturalness of synthesised speech – and thus enhance the communication experience – we apply the natural language processing tasks of part-of-speech tagging and chunking to the text in the synthesis process. We cover the South African languages of (South African) English, Afrikaans, isiXhosa, isiZulu and Sepedi. The part-of-speech tagging delivers positive results for most of the languages; however, the chunking does not give any improvement in its current form.

Index Terms: natural language processing, part-of-speech tagging, chunking, text-to-speech synthesis, prosody, human-computer interaction

1. Introduction

At the heart of the human condition lies communication. In our daily interaction, we rely on different aspects of communication to relate to one another – to speak and be spoken to, to hear and be heard, to understand and be understood. We do this very naturally via verbal speech and written text.

However, for some people, it is very difficult to communicate in a natural way, whether due to physical, emotional or cognitive challenges. The print-disabled individual cannot read a book in the conventional way, nor can the person with a speech impediment speak out loud normally.

Text-to-speech (TTS) synthesis can address these challenges to empower and equip the individuals to communicate. It can synthesise written text into audio to which print-disabled persons can listen. During the recently completed Lwazi III project [1], we engaged with the print-disabled community to roll out TTS voices in various South African languages. In the current EU-GBS project [2], we are working with the South African augmentative and alternative communication (AAC) community to give a mother tongue TTS voice to those who do not have the natural means to speak.

In order for TTS to be adopted as a sustainable communication solution, it is important that the voices produce natural-sounding synthesised speech. One of the biggest contributing factors to the naturalness of a TTS voice is prosody. Prosody includes word-level stress and tone, and phrase-level stress, intonation and breaks [3].

This paper builds on previous work [4] by applying natural language processing (NLP) on the text to be synthesised, in an attempt to inform the prosody of a TTS voice better. We expand the coverage of South African languages to (South African) English, Afrikaans, isiXhosa, isiZulu and Sepedi, each with bigger speech corpora. We perform not only part-of-speech (POS) tagging on the languages, but also chunking. We integrate the NLP into the frontend of the Speect TTS system [5], which uses the HTS backend [6] in turn.

The objective is to model the word-level stress and tone, and phrase-level stress, intonation and breaking behaviour in the original voice artist audio recordings by incorporating POS and chunking information, to be reproduced in the synthesised speech. We do not use any specific acoustic parameters, but rather the implicit modelling facility in the HTS framework, enabling the HTS training algorithm to extract the acoustic parameters automatically from the data [1][2].

In the next sections, we shall elaborate on the definitions of POS tagging and chunking and their experimental applications to TTS. We shall present and discuss the results, and conclude.

2. POS tagging

The first NLP task is the fundamental one of part-of-speech (POS) tagging of the words in the text. A POS tag is a linguistic category assigned to a word in a sentence based upon its morphological and syntactic – or morphosyntactic – behaviour. Words receive POS categories according to the affixes they take (morphological properties) and/or according to their relationship with neighbouring words (syntactic properties) [4][7].

Example POS categories common to many languages are noun, verb, adjective and adverb. Words are often ambiguous in their POS categories. The ambiguity is normally resolved by looking at the context of the word in the sentence. POS tagging is the automatic assignment and disambiguation of POS categories to words in electronic text. The machine learning algorithms used in this process are discussed in [8].

For the POS-related components of the experiments in this paper, we use the freely available HunPos tagger [9], a hidden Markov model (HMM)-based tagger that achieves 96.58% accuracy for English on Wall Street Journal (WSJ)-derived data from the Penn Treebank II corpus [10]. NWU CTexT has developed POS annotations for text data mined from the government domain for the 10 non-English South African languages [11]. Their POS tagging evaluation results using the HunPos tagger are summarised for the covered TTS languages, including the aforementioned English result, in Table 1.

POS tagging could potentially be useful to model prosody because word-level stress and tone are dependent on the POS

http://dx.doi.org/10.21437/Interspeech.2016-1040

category of the word. For example, in English and Afrikaans, nouns often carry stress on different syllables than verbs. In isiXhosa, isiZulu and Sepedi, the morphosyntactic information in the POS tag could help to distinguish which tone to use for similarly spelled words [12].

Furthermore, phrase-level stress and intonation require a structure of which POS and other syntactic information are building blocks [13]. Even a simple content-function word rule requires the POS of a word to categorise it. Phrase breaks could either be predicted from this syntax-based structure (see next section), which in turn requires POS tagging, or possibly implicitly from the POS tags themselves [4].

Table 1. *POS tagging performance.*

Language	Accuracy
English	96.58%
Afrikaans	95.71%
isiXhosa	84.18%
isiZulu	83.83%
Sepedi	96.00%

3. Chunking

The second NLP task is chunking. To define chunking, it is necessary to understand what syntax is. Syntax models the grammatical structure of a sentence, in other words the structural dependencies among the constituent words. It builds upon parts of speech by grouping words that are structurally related by their POS tags into phrases. Example phrases are noun phrases (headed by a noun) (NP), verb phrases (headed by a verb) (VP) and prepositional phrases (headed by a preposition) (PP). The structure is recursive and is typically represented as a tree [7][8].

However, automating syntax with machine learning requires great amounts of data. Chunking is an approximate technique that flattens the recursion of syntax to a single level, hence needing less data. It is consequently also referred to as shallow syntax. It typically takes a list of words as input, along with their corresponding POS tags, and produces a list of words as output, along with tags denoting their chunking types. The latter are NPs, VPs and PPs, amongst others, with distinctions between the beginning and the rest of the chunk, for example "B-NP" and "I-NP", respectively [2]. The machine learning algorithms employed by chunking are similar to those of POS tagging [8].

For the chunk-related components of the experiments in this paper, we use the CRF++ sequential labeller [14], an open source implementation of conditional random fields (CRF) that achieve an F-score of 0.9438 for English on WSJ-derived data from the CoNLL-2000 shared task [15]. NWU CTexT has developed chunk annotations on top of their POS data for the 10 non-English South African languages [16]. Their chunking evaluation results using the CRF++ sequential labeller are summarised for the covered TTS languages, including the aforementioned English result, in Table 2.

Even though the relationship between syntactic and prosodic phrasing is not always one to one [13], chunking could potentially model the prosodic effects that are more directly linked to syntax.

Table 2. *Chunking performance.*

Language	F-score
English	0.9438
Afrikaans	0.9517
isiXhosa	0.8545
isiZulu	0.9156
Sepedi	0.9755

4. Experiments

4.1. Common setup

We build the TTS voices from text and speech data using Speect and the HTS framework. For the English voice, we borrow text prompts from the CMU_ARCTIC speech synthesis databases [17] and books in the public domain. The Afrikaans voice also uses prompts from books and other writings in the public domain. For the voices in the other languages, we select text prompts from the same government domain data that NWU CTexT has collected. We record the voice artists reading the prompts out loud naturally, in a professional recording studio, each over a week. They are all female. The sizes of the text and speech corpora for the various languages are listed in Table 3.

Table 3. *Text and speech corpus statistics.*

Language	#Prompts	#Hours
English	4443	07h40m
Afrikaans	3878	06h50m
isiXhosa	1705	05h45m
isiZulu	1708	06h00m
Sepedi	2607	06h00m

In a series of experiments for each language, we progressively add the NLP information as HTS label features to the TTS voices and test whether it has an effect on the prosody or not. There are five resulting versions of the voices.

The baseline – with no NLP – uses only the default positional and counting features in the HTS labels, where punctuation delimits phrases broadly. It is termed "DEF".

The first POS version – on top of the baseline – adds POS categories of the previous, current and next word ("d1", "e1" and "f1"). It is termed "POS1".

The first chunking version – on top of the first POS version – replaces the broad punctuation-based phrase boundaries with more narrowly defined chunk boundaries as delimiters of phrases. It is termed "CHK1".

The second POS version – on top of the first POS version – reverts back to broad punctuation-based phrases, but emulates narrow intra-phrase behaviour by adding positional and counting features based on the higher content-function POS category of the current word ("e5", "e6", "e7" and "e8"). It is termed "POS2".

The second chunking version – on top of the second POS version – attempts to emulate narrow intra-phrase behaviour further by adding the current phrase-final chunk category in the stead of the ToBI end tone ("h5"). It is termed "CHK2".

Each synthetic voice with a new addition of NLP is compared to the previous synthetic voice without that addition, starting

with the baseline. This is done by determining which voice renders synthesised speech segments – from the text in a test set – that is closer to the original natural speech segments in the same set. The test set of each language comprises 100 randomly selected, grammatically conventional – simple and compound, flat and embedded – sentences.

The comparison of speech segments is made possible by the timing information that is inherently available in the synthesised speech, and obtained by aligning the original speech to the text. However, we compare at the word level, since word boundaries are more robust than phonetic boundaries in the automatic alignment of speech to text.

We calculate the distances between the synthesised and natural segments for the acoustic measures of duration, F0 and intensity, which have been shown to be correlates of prosody [18]. The distances for the latter two time-series are represented by their Euclidean distance-based dynamic time warping (DTW) costs.

We determine the statistical significance of the voice comparisons with McNemar's test, a chi-square test for paired sample data [19]. If the chi-square value is greater than or equal to 3.841, the synthetic voice with the most votes is significantly closer to the natural voice than the other synthetic voice. If the chi-square value is less than 3.841, the result is insignificant and the two synthetic voices can be said to be similar in closeness to the natural voice. The details and rationale of this approach are explained in [4].

The next subsections elaborate on the results for each language by means of tables. Each table lists the total test sample (word) allocations to the different voices for the three acoustic measures of duration, F0 and intensity. Columns 3 and 4 list the number of samples accredited to each voice. The last column in the table indicates the McNemar chi-square scores. Bold values highlight the statistically significant differences.

4.2. DEF-POS1 comparisons

The DEF-POS1 comparative test results are listed in Table 4.

Table 4. *Results of the DEF-POS1 comparisons.*

Language Measure	Total	DEF	POS1	Equal	Chi-square
English	1546				
Duration		**806**	695	45	8.135
F0		**876**	670	0	27.316
Intensity		**1069**	477	0	226.308
Afrikaans	1599				
Duration		699	**777**	123	4.069
F0		755	**844**	0	4.898
Intensity		806	793	0	0.098
isiXhosa	1233				
Duration		550	**670**	13	11.705
F0		261	**972**	0	409.416
Intensity		428	**805**	0	114.965
isiZulu	1157				
Duration		379	**774**	4	134.979
F0		210	**947**	0	468.826
Intensity		436	**721**	0	69.957
Sepedi	1881				
Duration		823	**979**	79	13.419
F0		419	**1462**	0	577.781
Intensity		658	**1223**	0	169.410

The Nguni family languages of isiXhosa and isiZulu, and the Sotho-Tswana family language of Sepedi show an improvement in prosody for all three measures (in terms of closeness), when the first version of POS category features are added. The F0 dimension is especially enhanced, which confirms the notion for these tonal languages that morphosyntactics and tone are linked.

The positive effects are less pronounced for the Germanic language of Afrikaans, albeit still significant. For Germanic English, the baseline without POS is actually closer to the original speech than the version with POS. However, this is not to say that morphosyntactics should be ruled out as an effective way to model the stress in these languages.

A possible explanation for the poor results of English could be a cross-domain drop in performance of the English POS tagger. It was trained on WSJ financial domain data, whereas the text sources for the TTS voice are more general fiction and non-fiction. Conversely, the isiXhosa, isiZulu and Sepedi POS taggers were trained on the very same government domain data from which the TTS text prompts were selected. Afrikaans is in the middle, with the POS tagger trained on the government domain data, while the TTS text sources are mainly general fiction and non-fiction.

4.3. POS1-CHK1 comparisons

The POS1-CHK1 comparative test results are listed in Table 5.

Table 5. *Results of the POS1-CHK1 comparisons.*

Language Measure	Total	POS1	CHK1	Equal	Chi-square
English	1546				
Duration		768	735	43	0.703
F0		698	**848**	0	14.457
Intensity		762	784	0	0.299
Afrikaans	1599				
Duration		**923**	633	43	53.863
F0		**945**	654	0	52.777
Intensity		**895**	704	0	22.696
isiXhosa	1233				
Duration		**685**	538	10	17.549
F0		**691**	542	0	17.885
Intensity		**832**	401	0	150.308
isiZulu	1157				
Duration		595	551	11	1.651
F0		**712**	445	0	61.385
Intensity		**619**	538	0	5.601
Sepedi	1881				
Duration		**984**	809	88	16.983
F0		**1107**	774	0	58.775
Intensity		**1022**	859	0	14.038

The first chunking version does not compare well against the first POS version for all languages, except English in the F0 dimension. However, taking into account that the baseline performs better than the first POS version for English, we test this particular chunking result by comparing it to the baseline as well. Indeed, the baseline outperforms again for all three measures significantly.

These results would suggest that syntax on its own – especially in the way that it is approximated by chunking – is not effective to predict prosodic structure, at least given the

amount of text and speech data for these TTS voices, and how they are processed in the HTS framework.

It would be prudent to investigate the syntax-prosody interface in more detail in the future. In particular, the work on data-driven automation of prosodic analysis could provide valuable insight. AuToBI [20] is a tool that was developed specifically for American English, but it might generalise to an extent to (or help bootstrap) South African English. Momel/INTSINT [21] is another tool that is completely unsupervised and language-agnostic; hence it could be used for the other South African languages. There has already been an initial effort in the work of [22].

4.4. POS1-POS2 comparisons

For this round of investigation though, the broad punctuation-based phrase boundaries of the first POS version turn out to be more robust than the narrowly defined chunk boundaries of the first chunking version. We revert back to the former and, nonetheless, try to model syntactic effects on prosody indirectly via the content-function word features in the HTS labels.

The POS1-POS2 comparative test results are listed in Table 6.

Table 6. *Results of the POS1-POS2 comparisons.*

Language Measure	Total	POS1	POS2	Equal	Chi-square
English	1546				
Duration		705	722	119	0.191
F0		739	807	0	2.947
Intensity		745	801	0	1.992
Afrikaans	1599				
Duration		743	726	130	0.185
F0		805	794	0	0.069
Intensity		792	807	0	0.131
isiXhosa	1233				
Duration		622	566	45	2.593
F0		607	626	0	0.278
Intensity		606	627	0	0.341
isiZulu	1157				
Duration		534	581	42	1.939
F0		572	585	0	0.135
Intensity		594	563	0	0.804
Sepedi	1881				
Duration		891	833	157	1.918
F0		953	928	0	0.319
Intensity		939	942	0	0.003

The differences across languages and measures are all statistically insignificant, indicating that these indirect features in the second POS version do not have a strong influence on prosody.

4.5. POS2-CHK2 comparisons

We make a final attempt at implicit prosodic modelling by approximating the ToBI end tone feature in the HTS labels with the category of the final chunk in the broad punctuation-based phrase. We compound it with the indirect content-function word features.

The POS2-CHK2 comparative test results are listed in Table 7.

Table 7. *Results of the POS2-CHK2 comparisons.*

Language Measure	Total	POS2	CHK2	Equal	Chi-square
English	1546				
Duration		758	707	81	1.741
F0		760	786	0	0.421
Intensity		**1086**	460	0	253.073
Afrikaans	1599				
Duration		**813**	686	100	10.675
F0		806	793	0	0.098
Intensity		**997**	602	0	97.330
isiXhosa	1233				
Duration		601	588	44	0.131
F0		618	615	0	0.005
Intensity		**773**	460	0	79.202
isiZulu	1157				
Duration		**602**	530	25	4.516
F0		**618**	539	0	5.326
Intensity		**735**	422	0	84.405
Sepedi	1881				
Duration		882	857	142	0.345
F0		953	928	0	0.319
Intensity		771	**1110**	0	60.916

The results are either insignificant, or they are biased towards the second POS version without the ToBI-approximated feature. The second chunking version shows a single exception with Sepedi in the intensity dimension; however, the reason for this outlying behaviour is unknown.

The non-effects and adverse effects of the implicit features in the second POS and chunking versions seem to strengthen the argument for more explicit prosodic modelling.

5. Conclusions

We have shown that POS tagging can be applied effectively to TTS in the HTS framework to improve prosody for a subset of South African languages. Further syntactic analysis, in the form of chunking, is not yet successful.

However, a roadmap has been drawn for future work to incorporate more explicit prosodic analysis by way of AuToBI and Momel/INTSINT. In TTS, the problem still remains to link the prosodic analysis back to the text, since the prosody can be predicted from the text alone at synthesis time (barring direct specifications, for example via SSML). To improve upon the results of [22], the answer might lie with more comprehensive syntactic features to bridge the gap to the INTSINT features in the syntax-prosody interface.

Notwithstanding the attempts in the lab to improve the TTS voices, it is important to remain cognisant of end-user needs and preferences. During the perceptual evaluations of the Lwazi III project, some print-disabled end-users already voiced satisfaction with the baseline voices, mostly because of the opportunity to read in their own language for the first time. Others did set higher standards for intelligibility and naturalness. This shows that we can deploy our TTS voices and change lives, even if the voice quality is not yet perfect.

6. Acknowledgements

We would like to thank NWU CTexT for providing us with the South African text and NLP-annotated data.

7. References

[1] N. Titmus, G. I. Schlünz, J. A. Louw, A. Moodley, T. Reid, and K. Calteaux. Final Report: Lwazi III Operational Deployment of Indigenous Text-to-Speech Systems. CSIR Meraka Institute, Pretoria, South Africa, 2016.

[2] F. de Wet, K. Calteaux, J. A. Louw, J. Badenhorst, A. Moodley, C. Moors, G. I. Schlünz, and N. Titmus. Annual Progress Report: EU-GBS HLT speech-enabled service delivery platform. CSIR Meraka Institute, Pretoria, South Africa, 2016.

[3] P. Taylor. Text-to-Speech Synthesis. First edition, Cambridge University Press, 2009.

[4] G. I. Schlünz, E. Barnard, and G. B. van Huyssteen. Part-of-speech effects on text-to-speech synthesis. In Proceedings of the 21st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), pages 257-262, Stellenbosch, South Africa, 2010.

[5] J. A. Louw. Speect: A multilingual text-to-speech system. In Proceedings of the 19th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), pages 165-168, 2008.

[6] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda. The HMM-based speech synthesis system version 2.0. In Proceedings of ISCA SSW6, pages 294-299, 2007.

[7] D. Jurafsky and J. H. Martin. Speech and Language Processing. Second edition, Pearson Education, 2009.

[8] G. I. Schlünz. The effects of part-of-speech tagging on text-to-speech synthesis for resource scarce languages. MSc dissertation, North-West University, Potchefstroom, South Africa, 2010.

[9] HunPos. https://code.google.com/archive/p/hunpos/ . Accessed 22 March 2016.

[10] P. Halácsy, A. Kornai, and C. Oravecz. HunPos: an open source trigram tagger. In Proceedings of the 45th annual meeting of the Association for Computational Linguistics (ACL) on interactive poster and demonstration sessions, pages 209-212, 2007.

[11] E. R. Eiselen, and M. J. Puttkammer. Developing text resources for ten South African languages. In Proceedings of the 9th International Conference on Language Resources and Evaluation, pages 3698-3703, Reykjavik, Iceland, 2014.

[12] S. Zerbian and E. Barnard. Word-level prosody in Sotho-Tswana. Speech Prosody, volume 100861, pages 1-4, 2010.

[13] M. Steedman. Information structure and the syntax-phonology interface. Linguistic inquiry 31.4, pages 649-689, 2000.

[14] CRF++. https://taku910.github.io/crfpp/ . Accessed 22 March 2016.

[15] F. Sha, and F. Pereira. Shallow parsing with conditional random fields. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pages 134-141, 2003.

[16] E. R. Eiselen. South African language resources: phrase chunkers. In Proceedings of the 10th International Conference on Language Resources and Evaluation, Portroz, Slovenia, 2016 (In Press).

[17] CMU_ARCTIC data. http://festvox.org/cmu_arctic/ . Accessed 22 March 2016.

[18] A. Waibel. Prosody and Speech Recognition. First edition, London: Pitman Publishing, 1988.

[19] S. Boslaugh, and P.A. Watters. Statistics in a nutshell. First edition, O'Reilly Media, Inc., 2008.

[20] A. Rosenberg. AutoBI-a tool for automatic toBI annotation. In Proceedings of Interspeech 2010, pages 146-149, 2010.

[21] D. J. Hirst. A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. In Proceedings of the XVIth International Conference of Phonetic Sciences, volume 12331236, 2007.

[22] J. A. Louw and E. Barnard. Automatic intonation modeling with INTSINT. In Proceedings of the 15th Pattern Recognition Association of South Africa (PRASA), pages 107-111, South Africa, 2004.