INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Nativization of foreign names in TTS for automatic reading of world news in Swahili

Joseph Mendelson[1], *Pilar Oplustil*[2], *Oliver Watts*[2], *Simon King*[2]

[1]KTH Royal Institure of Technology, Sweden
[2]Centre for Speech Technology Research, University of Edinburgh, Edinburgh, EH8 9AB, UK
josephme@kth.se, psoplust@uc.cl, owatts@staffmail.ed.ac.uk, Simon.King@ed.ac.uk

Abstract

When a text-to-speech (TTS) system is required to speak world news, a large fraction of the words to be spoken will be proper names originating in a wide variety of languages. Phonetization of these names based on target language letter-to-sound rules will typically be inadequate. This is detrimental not only during synthesis, when inappropriate phone sequences are produced, but also during training, if the system is trained on data from the same domain. This is because poor phonetization during forced alignment based on hidden Markov models can pollute the whole model set, resulting in degraded alignment even of normal target-language words. This paper presents four techniques designed to address this issue in the context of a Swahili TTS system: automatic transcription of proper names based on a lexicon from a better-resourced language; the addition of a parallel phone set and special part-of-speech tag exclusively dedicated to proper names; a manually-crafted phone mapping which allows substitutions for potentially more accurate phones in proper names during forced alignment; the addition in proper names of a grapheme-derived frame-level feature, supplementing the standard phonetic inputs to the acoustic model. We present results from objective and subjective evaluations of systems built using these four techniques.

Index Terms: speech synthesis, TTS, Under-resourced languages, code-switching, multi-lingual speech synthesis, text processing

1. Introduction

Swahili (natively known as Kiswahili) is spoken by more than 100 million people across East and Central Africa, but is still considered under-resourced (or low resource) by the speech and language technology community [1]. In collaboration with a UK-based international news organisation, we attempted to build a production-quality Swahili text-to-speech (TTS) voice that could be used to re-broadcast news stories. Through this organisation, we obtained access to a professional quality voice talent (a Swahili-speaking, on-air journalist) for data collection, and Swahili text corpora from their international news archive. Such international news corpora inherently contain a large percentage of proper names, originating in a vast array of languages. The most common examples were politicians, athletes, and figures from popular culture. This created pronunciation challenges for the speaker, as well as technical challenges throughout the TTS-building pipeline, which we address in this paper.

The orthography of Swahili is highly transparent: the grapheme-phoneme relation is usually one-to-one, and as such letter-to-sound (LTS) rules are an appropriate approach to obtain phonetic transcriptions. However, for non-Swahili proper names, the speaker of our database employs code switching: changing to a another language while speaking [2]. These code-switched proper names are exceptions to the LTS rules because the speaker modifies his pronunciation *towards* the language of origin of the proper name. He does not necessarily *achieve* the pronunciation from the language of origin, and neither is he consistent in his attempts. News organisations provide pronunciation guidelines for their journalists, but in practice speakers tend to 'blend' their native accent with the target pronunciation [3]. A possible solution for TTS might be to switch LTS rules, but it is extremely difficult to automatically determine the language of a single foreign word in the middle of a sentence [4, 5]. In any case, proper names often do not follow the pronunciation rules of their language of origin.

Because the goal is to create a usable synthetic voice for international news, the challenge of synthesising international foreign names with accurate pronunciations must be addressed. In our corpus, they constitute approximately 18% of text. It is reasonable to assume that future news text will have a similar proportion, including many previously-unseen proper names. An automated solution is desirable.

The work presented here starts with a baseline system built using a basic Swahili LTS rule set, including syllabification and stress marking, based on published accounts of the linguistic characteristics of Swahili [6, 7, 8]. When using only Swahili LTS rules, the large quantity of foreign names in our corpus led to many inaccurate phonetic transcriptions. This substantially affected forced alignment of the data. We use Deep Neural Network (DNN) speech synthesis, and such errors in forced alignment lead to *general* degradation in output quality.

We describe four approaches – each building incrementally on one or more of the previous techniques – to improve foreign proper name transcriptions, as well as overall objective and perceptual performance:

- Utilising a UK Received Pronunciation lexicon and LTS rules (Unilex RP [9]) to automatically generate an addendum to the lexicon, for proper names.
- Adding custom phones to this addendum to differentiate them from the regular, Swahili LTS-derived phones, as well as adding a part of speech (POS) tag for proper names to these words, which is employed as a frame-level input feature for the DNN.
- Allowing certain of the custom phones to substitute for similar potential phones during forced alignment.
- Adding a frame-level feature that indicates whether a word's transcription originates from LTS rules or the addendum, or if it was an acronym. If it comes from the addendum, grapheme features are also added.

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-1398

2. Related Work

Llitjos & Black [5] investigate whether knowledge of a proper name's native language improves its pronunciation accuracy in a US English TTS system. They obtain the probability of an input word under each of 26 letter-trigram language models, all trained on text from different languages. Features derived from these 26 scores are used as extra input features when learning a decision tree for phonetization. This method improved listening test scores by 17%. Meron [10] takes an approach to reducing the effect of transcription errors for British names which is similar to one used in the current work, developing context sensitive rules to find such errors in order to correct them and clean the database (although this is done in a manual, iterative fashion). A special subjective evaluation was carried out, asking people to rate proper name pronunciation as 'correct', 'acceptable', 'wrong', 'intermediate' or 'bizarre'; their method achieved on average more "acceptable" judgements. This shows the significance of cleaning the data and isolating non-target-language data from proper names, as well as the fact that the evaluation of proper name synthesis is an open problem. Spiegel [11] describes a research program aimed at solving the proper name problem in the context of US telephone companies. Their system uses a small dictionary of exceptions, an ethnographic classification module that tries to determine the language of origin, and hand-tuned LTS rules. The LTS rules were improved by doing large-scale polling to determine acceptable pronunciations for proper names.

Research on how to handle the pronunciation of proper names has also been done in the context of speech recognition. For example, Reveil et al. [3] developed multilingual acoustic models combined with grapheme to phoneme transcriptions, plus phoneme to phoneme (P2P) transcriptions. P2P models are trained to map phones from one language to phones of a different one, as speakers do when they transfer pronunciations from their native language to the target language of the proper name they want to pronounce.

3. Methodology

In a pilot experiment (phase 1), we collected half of the database, after which we inspected the data, built the first voice (V1), and made adjustments to the methodology. Informal listening at the end of phase 1 suggested that we had achieved a reasonably high quality output, with the notable exception of proper names. In phase 2 we collected the remaining half of the data, built the baseline voice (V2), and then proceeded to focus on the the proper names challenge (voices V3, V4, V5, and V6). We used the Festival framework [12] to build our front end for Swahili, for use in all experiments. We derived DNN inputs in a standard way by using HTS-style context-dependent phone labels as an intermediate representation, converting to binary, upsampling this to acoustic frame rate, and adding sub-phonetic positional features.

All DNNs were built with the OpenSource Merlin toolkit [13] and the same DNN architectures was used in all voices. Acoustic model: 4 feedforward layers of 1024 neurons with *tanh* activation function, plus 2 simplified long short-term memory (SLSTM, [14]) layers of 512 neurons. Duration model: 2 feedforward *tanh* layers of 512 neurons and 2 SLSTM layers of 512 neurons.

3.1. Text processing modules

The LTS rules used and the basic phone set were derived from a combination of previous work in ASR [15], web resources [16], and our own linguistic knowledge. This included some limited digit normalisation, basic syllabification, and a simple universal stress marker for the second-to-last syllable.

3.2. The text corpus and speech database

All experiments used the same text corpus of approx. 40,000 Swahili sentences retrieved from a television and radio script archives, plus web content. From this corpus, we extracted a sub-corpus of 2029 sentences, using an algorithm to rank sentences according to the relative frequency of their diphones (as generated by our LTS rules). This sub-corpus was recorded by the voice talent in two phases, resulting in a final speech database of approx. 5 hours and diphone coverage of 71% (phoneset size is 44).

All evaluations were performed on a held-out test set of 10 sentences, hand-chosen for their high percentage of proper names from many languages (e.g. Spanish, Korean, Chinese, Hungarian, Dutch, English, Bengali, Turkish and Italian).

3.3. The Baseline Voice (V2)

The baseline was built using the full speech database and with LTS rules. It was clear that incorrect proper name transcriptions predicted by our LTS rules were not only a problem during synthesis, but that they also negatively affected forced alignment of the entire database, and therefore the trained DNN as well, reducing the naturalness of *all* speech generated. The benefit which would normally be expected from doubling the size of the speech database (from V1 to V2) was being undermined by a doubling in quantity of poorly-transcribed proper names.

3.4. The Unilex RP Lexicon and LTS: a Semi-Automatic Addendum (V3)

Careful inspection of the database revealed that the speaker tended towards a UK Received Pronunciation (RP) accent when pronouncing names of English-language origin. Analysing the Unilex RP lexicon and LTS [9] we found that, in addition to English names, many common European language names, and even some Asian ones, would be transcribed close to our speaker's pronunciation. We used Unilex RP to semi-automatically build an addendum to our front-end for proper names. A list of likely proper names was isolated by selecting non-sentence-initial words that were capitalised. To remove Swahili names from that list, a trigram letter language model for Swahili was trained, based on the non-capitalized words that remained in the corpus after the likely proper names had been removed. We scored the candidate proper names with this model, and all words scoring above a manually-tuned threshold were regarded as Swahili and removed from the list, resulting in a list of mostly non-Swahili names from a multitude of languages. Unilex RP was used to generate phonetic transcriptions for items in the list. The Unilex RP phone set was simply mapped to the Swahili phone set using hand-crafted rules, adding only 2 new phones.

3.5. Proper Name Phone Set, POS Tag, and Reduced Addendum (V4)

Building on V3, we attempted to further improve forced alignment, by adding a set of distinct phones for the proper names in the addendum. The intention was to improve the HMMs used for forced alignment, by separating the phones used for proper names from those used for regular Swahili words. This would potentially improve proper name alignment, but more importantly also the alignment of all other text. We added a special POS tag for proper names as a frame-level feature, to allow the DNN to learn different input-output mappings for each phone set.

A final informal inspection of the V3 lexicon addendum and the text corpus indicated that some Swahili words appeared in the addendum and should be removed and processed instead by the Swahili LTS rules: in many cases, these same words appeared elsewhere in the text, but in non-capitalised form, indicating they may not actually be proper names at all, but also that they are likely to be Swahili.

3.6. Phone Substitutions During Forced Alignment (V5)

While the lexicon addendum and associated techniques in V4 were a step in the right direction, it became clear that new errors were arising for some names. This is not surprising, given the wide range of languages of origin in our data: no single-language lexicon/LTS would be able to transcribe them all accurately. To address this, we inspected the recordings and we designed a set of phone substitutions based on our analysis of typical grapheme-phoneme alternative realisations when using the Latin alphabet for multiple languages. These substitutions were permitted only within proper name transcriptions. During forced alignment, this allowed incorrect transcriptions to be 'repaired' based on acoustic evidence: for example, the grapheme <ch> is transcribed as an affricate by the Unilex LTS rules, but in an Italian name a [k] would be more appropriate.

3.7. LTS/Addendum/Acronym Discrimination and Mapping Phones to Graphemes (V6)

There will always be some inconsistent mappings between proper name phones and graphemes. For example, for the names "Kun" and "Khan" in the addendum, after mapping to Swahili phones both words have the transcription [k a n]. However, "Kun" (the nickname of an Argentinian football player) should be pronounced closer to [k u n].

To address the general problem of the phonetic transcription failing to capture these subtleties, we added two input features to the DNN. The first indicates whether the transcription was generated using Swahili LTS, or the addendum, or corresponds to an acronym (we currently do not have a special normalisation module for acronyms).

The second feature indicates, for the phones of words in the lexicon addendum, which graphemes were most likely to be aligned with them. In the "Kun" vs. "Khan" example, this feature would indicate whether the phone [a] arose from a <u> grapheme or an <a> grapheme. Implementing this feature requires recovery of the grapheme-phoneme alignment, for which we used an algorithm based on Minimum Edit distance. The new feature was encoded as a one-of-K, with K covering all single- and multi-graphemes that a single phone can align to (K=300 for our data). For example, Table 1 shows the alignment of graphemes and phones for the place name "Marseille".

Table 1: *Phones and grapheme alignment for the proper name "Marseille".*

Phones	m	a	s	e	i
Graphemes	m	ar	s	e	ille

4. Results

We first report objective scores for all of our models: Table 2. Compared to the baseline, the techniques applied in V4 lead to the most improvement in the acoustic model, and those in V6 generated the most improvement in the duration model.

Table 2: *Objective scores: MCD and F0 RMSE for acoustic model and RMSE (frames per phone; 5ms frame shift) for duration model.*

System version	Acoustic MCD (dB)	Acoustic F0 RMSE (Hz)	Duration RMSE
V2	5.234	31.765	10.342
V3	5.195	32.366	11.398
V4	**5.120**	**30.763**	9.992
V5	5.253	32.615	9.915
V6	5.134	32.305	**7.417**

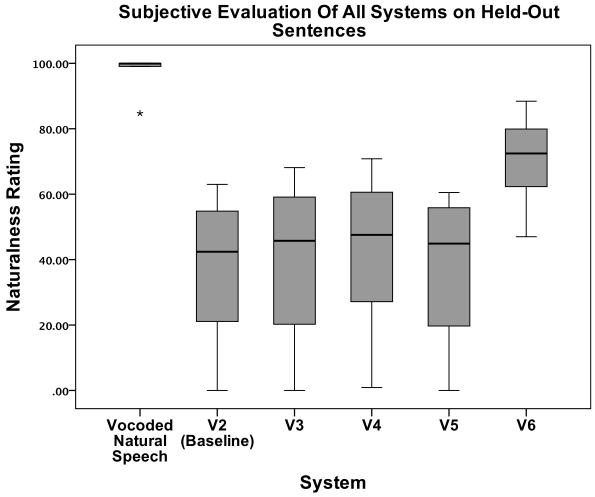

Figure 1: *Boxplot comparing listeners' naturalness ratings from a MUSHRA-style evaluation of complete Swahili sentences containing foreign proper names. Medians are shown as black horizontal bars.*

Subjective listening tests were conducted online, and 9 fluent Swahili-speaking participants were recruited via the Prolific Academic crowd-sourcing platform [17]. A MUSHRA-style test was employed, to compare 6 systems: the V2 Baseline, V3-V6, and Vocoded Natural Speech (as the hidden reference). Each MUSHRA screen presented 6 versions of a single sentence, and there were 10 screens (one for each of the 10 held-out test sentences). Subjects were instructed to rate each stimulus for naturalness on a scale from 0-100. Participants were given the following instructions: "By 'naturalness', we mean: 'sounds like a human voice'. On our Scale of 0–100, 100 would be the *most* natural (sounds like a human), and 0 would be the *least* natural (sounds like a robot)".

Each listener was presented with the 10 sentences (MUSHRA screens) in a different randomised order; the left–right ordering of the 6 stimuli within each MUSHRA screen was randomised per-screen. Results are presented in Figure 1. System V6 was perceived as significantly more natural than all other systems (Wilcoxon Signed Rank test with Bonferroni correction; $p < .05$).

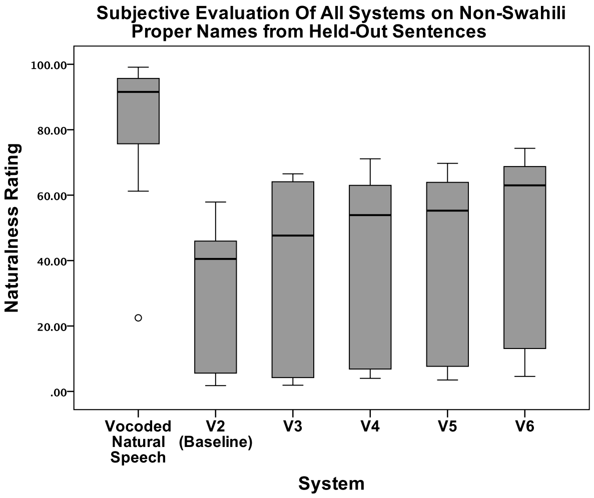

Figure 2: *Boxplot comparing listeners' naturalness ratings from a MUSHRA-style evaluation of isolated proper names of non-Swahili origin. Medians are shown by black horizontal bars.*

The same listeners evaluated 10 proper names from the test sentences, originating from 10 non-Swahili languages, synthesised in isolation, again in a MUSHRA-style comparative format. The Vocoded Natural Speech stimulus, which was used as a hidden reference, was excised from the naturally-spoken utterance used in the other part of the listening test.

Presentation order was randomised in the same way as the other part of the listening test. Results are presented in Figure 2. V6 was slightly preferred, although not significantly. Since there is no gold-standard for correct pronunciation of proper names, this higher between-listener variability is perhaps to be expected.

In the results presented in Figures 1 and 2, one participant has been discarded due to inability to rate Vocoded Natural Speech as the most natural, leaving responses from 8 listeners.

5. Discussion

5.1. Comparing the systems built

We have developed and tested a range of approaches for improving the transcription, forced alignment and synthesis of proper names. The results show a general improvement, both objectively and subjectively for our Swahili TTS systems over the baseline. Perhaps the most noteworthy result comes from the frame-level grapheme-derived features in V6. Whilst these did not improve the acoustic model, they improved duration modelling very substantially which is presumed to be the reason for listener's clear preference for this system when synthesising sentences (Figure 1).

The phone substitution technique of V5 did not have much effect compared to V4, but only one set of possible substitutions was tested.

5.2. Proper names

As a general conclusion, separating proper names from the main text and processing them separately seems to be very effective. Unilex RP was helpful for generating a transcription closer to our speaker's pronunciation for the majority of proper names, although there were exceptions where Swahili LTS rules were more accurate. This suggests that it is important to carefully filter words, which we achieved with a character-based trigram model.

Our effective use of Unilex RP highlights the utility of mixing resources from different languages to build TTS systems for languages that have low resources, particularly in a domain such as world news.

5.3. Low resource languages

This work presented us with many of the challenges typical of working with low resource languages. None of the authors speaks Swahili and we had limited access to native speakers. The text corpus contained many misspellings that went unnoticed until the voice talent was asked to record the sub-corpus. The building of the front-end was informed primarily by external resources, without help from native speakers. Unsurprisingly, listening tests participants were somewhat difficult to find, whether locally or via web-based crowd-sourcing platforms.

5.4. Future work

We would like to explore multilingual phone-level recognition approaches (such as [18]), as a potential means of generating accurate transcriptions for foreign words. The relatively simple method we used to identify proper names could be improved through various natural language processing techniques, such as named entity recognition or full part-of-speech tagging. We would like to further validate the techniques explored here by applying them to another language in the news domain, or for TTS in another domain where proper names are particularly relevant.

6. Conclusion

There are unique challenges encountered when working on TTS for any low resource language. In our particular circumstance of working with an international news organisation, we were faced with the issue of a very high percentage of foreign proper names originating from unspecified languages. LTS rules for the primary language proved inadequate for these names. They produced erroneous transcriptions which degraded forced alignment, DNN acoustic and duration models, and ultimately the synthetic speech. The approaches described in this paper can alleviate this problem to a considerable degree.

7. Acknowledgements

The research leading to these results was partially supported by EPSRC Standard Research Grant EP/P011586/1, *Speech Synthesis for Spoken Content Production (SCRIPT).*

8. References

[1] L. Besacier, E. Barnard, A. Karpov, and T. Schultz. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100, January 2014.

[2] L. Yu, W. He, W. Chien, and Y. Tseng. Identification of code-switched sentences and words using language modeling approaches. *Mathematical Problems in Engineering*, vol.2013(Article ID 898714), 2013.

[3] B. Réveil, J. P. Martens, and H. van den Heuvel. Improving proper name recognition by adding automatically learned pronunciation

variants to the lexicon. In *7th Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2149–2154, Paris, France, 2010.

[4] Y. Chen, J. You, M. Chu, Y. Zhao, and J. Wang. Identifying language origin of person names with n-grams of different units. In *Proc. ICASSP*, Toulouse, France, 2006.

[5] A. F. Llitjos and A. W. Black. Knowledge of language origin improves pronunciation accuracy of proper names. In *Proc. Interspeech*, pages 1919–1922, Aalborg, Denmark, 2001.

[6] Polomé E. *Swahili language handbook.* Center for Applied Linguistics, 1967.

[7] M. Gakuru, F. Kang'ethe, and K Ngugi. Some essential features in developing a text to speech system in Kiswahili. In *Proc. LLSTI workshop*, Lisbon, Portugal, 2004.

[8] W. Ngugi, P. Okelo-Odongo, and W. Wagacha. Swahili text to speech system. *African Journal of Science and Technology (AJST) Science and Engineering Series*, 6:80–89, 2004.

[9] S. Fitt and S. Isard. Synthesis of regional english using a keyword lexicon. In *Proc. Eurospeech-99*.

[10] J. Meron. Using rules to improve letter to sound conversion of names. In *Proceedings of 2002 IEEE Workshop on Speech Synthesis*, pages 59–62, 2002.

[11] M. F. Spiegel. Proper name pronunciations for speech technology applications. *International journal of speech technology*, 6(4):419–427, 2003.

[12] R. Clark, K. Richmond, and S. King. Multisyn: Open-domain unit selection for the festival speech synthesis system. *Speech Communication*, 49:317–330, 2007.

[13] Z. Wu, O. Watts, and S. King. Merlin: An open source neural network speech synthesis system. In *Proc. SSW*, Sunnyvale, USA, 2016.

[14] Z. Wu and S. King. Investigating gated recurrent networks for speech synthesis. In *Proc. ICASSP*, pages 5140–5144, Shanghai, China, 2016.

[15] Hadrien Gelas, Laurent Besacier, and Francois Pellegrino. Developments of Swahili resources for an automatic speech recognition system. In *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape-Town, Afrique Du Sud, 2012.

[16] S. Ager. Omniglot - writing systems and languages of the world. consulted on august 2016. *www.omniglot.com*, 2015.

[17] Prolific. Prolific – find participants fast. https://prolific.ac/. Accessed: 2017-3-10.

[18] J. Köhler. Multilingual phone models for vocabulary-independent speech recognition tasks. *Speech Communication*, pages 21–30, 2001.