HLTs in Second Language Learning – Can we optimise TTS technology to be a better teaching tool?

Lehlohonolo Mohasi, Mqhele E. Dlodlo, *Members IEEE*

University of Cape Town
Department of Electrical Engineering
Private Bag X3
Rondebosch, Cape Town
South Africa, 7700

lmohasi@crg.ee.uct.ac.za; Mqhele.Dlodlo@uct.ac.za

Abstract – This paper investigates the possibility of optimising text-to-speech (TTS) technology for a more efficient use in second language learning. We believe that a high quality TTS system first has to be developed before this technology can be optimally used in second language learning. In reviewing the different techniques and methods by different researchers in TTS technology with the aim of advancing TTS speech output, we also try our hand by directing our focus on improving prosody of a TTS system. Since language learning is best done through interaction, we propose an interactive TTS system through dialogue. Our attempt therefore, at improving prosody, will be done in relation to a dialogue system. Another proposal is the integration of linguistic distance measure with TTS. We believe this will be more applicable to the source and target languages of interest being studied.

We finally discuss the different methods used for evaluating TTS systems in general, and evaluation of TTS performance in language learning. We show that people use different methods and metrics depending on what aspect or criterion they want to measure. An overall agreement has not been reached as yet to the technique or metric which gives a true assistive measure of the technology. We conclude with a remark that an interactive TTS system with prosody enhanced is one good step towards an optimal teaching tool.

Index Terms – high quality TTS, linguistic distance measure, prosody, second language learning

I. Introduction

Human Language Technologies (HLTs) have progressed significantly and are widely used in many applications, language teaching/learning being one of them. Technologies such as automatic speech recognition (ASR) and text-to-speech (TTS) are commonly used in computer-aided language learning (CALL) applications, despite their limitations. These technologies, when used in an integrated system, form an interactive interface in which learners can interact with a computer in a form of a dialogue.

This work was supported in part by Telkom SA, Nokia Siemens, the National Research Foundation and the Department of Trade and Industry.

Although these CALL applications have been widely embraced, they still do not meet all the expectations of the learners. Our belief therefore, is that each technology needs to be highly developed in trying to meet learners' expectations. The focus of this paper thus, is to find a way of optimising TTS technology in order to make it a better teaching/learning tool for second language (L2) learners.

Since the goal is to reach an optimal level of TTS technology, we start off by investigating features which researchers believe play a role in a high quality TTS system.

II. What is a high quality TTS system?

TTS is one technology which has been vigorously researched though its study is ongoing as a high quality mark for speech output has not been reached as yet. The first success was a TTS system which sounded robotic [1]. Various approaches in speech synthesis have been tried and tested in order to develop a better sounding system than the previous ones. These approaches include rule-based formant synthesis, concatenative diphone synthesis, and concatenative unit selection synthesis [1], with the latter being of more interest as it produces a more natural sounding voice output.

Unit selection TTS synthesizes speech by concatenating selected units from a database of recorded speech. A unit selection algorithm selects a sequence of speech units “that best matches the targets (the desired characteristics as determined by the TTS front end) and also that join together most smoothly” [2]. As there is little signal processing performed on the resulting speech signal, minimal distortion is experienced. Speech output from a unit selection synthesis is more natural sounding than the other synthesis techniques.

According to Syrdal and Kim [2], even though unit selection produces a natural sounding speech, its prosody is more difficult to control. They suggest use of speech acts which they believe will provide more prosodic control of TTS. Their notion is that TTS control via speech acts will be at a more accessible level than that requiring specialised linguistic knowledge. Our aim is to verify this hypothesis vigorously and hope to obtain a more qualitative TTS speech output.

A. Basic TTS factors

In almost every project or system, the basics have to be taken care of before focusing on the major challenges. Text-to-speech technology is one such system which, according to [3], has two key elements that should first be considered in order to maximise the TTS output. These factors, if not tidied up, will hinder optimal development of a high quality TTS as they have control over input text. The two factors are context and text formatting.

Context involves syntactic, semantic and pragmatic analyses. Syntactic analysis is using the structure and order of the words, semantic analysis involves using the meaning of the words, and pragmatic analysis is the use of the real-world context and tone of the text.

Text formatting, on the other hand, is composed of three areas which frequently interact with the ability of a TTS system to carry out standard text analysis. These are punctuation, spelling and use of case.

Use of punctuation is critical to achieving the best quality TTS output. Inserting pauses intelligently into an unpunctuated string of text is one of the liveliest areas of speech synthesis research [3] – and it is not successful thus far.

Spell-checkers for TTS systems have not been developed as yet; therefore, correct spelling is essential for proper speech output. TTS systems read words precisely as they have been entered.

Text-to-speech systems are 'case sensitive' in that they rely on case to make decisions about how words should be pronounced. Use of case hence, affects the system's ability to process text accurately. Use of words in small case and upper case (in one utterance) causes issues for most TTS systems [3].

Once these basic features have been tackled, the problematic ones can be attended to. Researchers in speech technology have used different attempts to embark upon these 'higher level' TTS factors, as briefly explained below.

B. Attempts at advancing TTS technology

Research into making TTS systems sound natural has been tackled by different researchers in different ways. For instance, the paper written by Kim et al [4] describes how they reduced phone label errors in TTS voice building by means of modelling speaker pronunciation variants. They created speaker-dependent pronunciation lexicons for automatic speech labelling of their TTS voice databases. This effort helped eliminate many pronunciation errors that resulted from mismatches between lexical pronunciations and how the speaker actually pronounced a word. Thus, they managed to produce a TTS system of better speech output quality.

Conkie et al [5], on the other hand, refined existing pre-selection by adding multiple phone sets to the list of features considered. This led to a better database usage plus significantly increased synthesis quality [5]. In comparing their work with what had been done by Black and Taylor [6], though both groups' aim was to improve on existing synthesis, and reduce the number of low-scoring utterances, Black and Taylor had inconclusive results. Black and Taylor use decision trees, with acoustic differences as a similarity measure, whereas Conkie et al rely on finer context-based phone distinctions in conjunction with pre-selection optimization.

On the whole, Conkie et al had favourable results which led to the discovery that having multiple phone sets allows flexibility in the construction of the unit selection in general. This method, as they further point out, is also language independent in the sense that one is free to add new features they deem appropriate for a particular language.

From these brief reports, and many others not mentioned in this paper, it is obvious that no single method or technique is the best (up to now) in advancing TTS technology.

C. What is the best way forward?

Despite these developments, some of which produce state-of-the-art TTS systems, TTS technology still has not reached robustness as per evaluation by human subjects [7]. It still has limitations which logically do not make it the best teaching tool for language learners. We believe that the best approach in this sense is to develop a high quality TTS system, with as few limitations as possible. But then, we need to know what is meant by a high quality TTS system.

Our understanding in relation to this dilemma is that a high quality TTS system should sound natural like a human being (this includes understandability and intelligibility), be flexible in language identification and code-switching, and be able to express emotions (prosody). In short, it should be able to read and 'act' the text it is synthesising, much the same way a human reader would. It should not be dull, but it should be pleasant to listen to. Of interest in this paper, are language identification and prosody, which we will talk about.

Since a human being has an ability to identify different languages, the same is expected of TTS systems. This requires linguistic knowledge of the language in question in order to be able to set phonetics, letter-to-sound rules, etc. This aspect has been widely researched as TTS systems in a variety of languages are available, though we could say it is only at a half-way mark. The reason of saying this is because we need one TTS engine which can switch between languages, the same as one human brain switching between languages. The TTS systems which have been developed so far are monolingual.[1] These systems, though most of them sound natural, are lacking in terms of prosody.

Prosody is one other aspect which is still evading researchers and needs further investigation. Prosody includes aspects such as tone/intonation, stress, pitch, duration and emotional expression. As Syrdal and Kim indicated in their paper [2], "unit selection synthesis results in more natural sounding speech than the other synthesis techniques, but its prosody is more difficult to control." One is of the impression that an improvement of one TTS aspect is at a cost of another.

[1] The authors are not aware of a bilingual TTS system which uses one engine to synthesise more than one language simultaneously.

III. Optimising TTS technology for second language learning

Since naturalness (speech output sounding like a human being) by TTS has almost been reached, we decided to focus on other two aspects (language identification/distance measurement and prosody) which we believe if fully developed, will play a major role in the optimisation of TTS technology for language teaching and learning.

A. Language Distance Measurement

As the aim is to optimise TTS technology for language learning, the aspect of language identification in this part will focus on a 'novel' feature, language distance measurement. Linguistic distance is the extent to which languages differ from each other. We propose integrating this linguistic notion into TTS technology in order to advance language learning through technology use.

As indicated by Zulu [8], researchers spend countless hours converting and adapting speech technology systems to support different languages. Zulu further mentions that linguistic distances enable researchers to use resources of a source language to train models for a target language. This concept can consequently be used to train TTS models as well. If successful, this may prove to be "invaluable for the accelerated and efficient design and implementation of multilingual speech technology systems", text-to-speech included [8].

Approaches to linguistic distance measures are text-based and acoustic-based. The text-based approach involves orthographic and phonetic transcriptions, whereas acoustic based includes use of formants, phonetic, MFCC, etc.

Although the concept of language distance measure is well-known among linguists, the prevailing view is that is cannot be measured [9]. Researchers in speech technology are ploughing their way into the matter as well, and hopefully a scalar measure can be developed for linguistic distance.

Languages are complex and differ in vocabulary, grammar, written form, syntax, and myriad other characteristics. This makes for difficulty in the construction of measures of linguistic distance [9]. The distance between two languages may also depend on whether they are in the written or spoken form. For example, the written form of Chinese does not vary among the regions of China, but the languages differ sharply.

Knowledge of linguistic distance may be invaluable for understanding differences across groups in the acquisition of target language skills learners, or "the linguistic issues facing indigenous linguistic minorities (e.g. indigenous languages speaking people in Africa), and the complexity of adaptation in multilingual societies" [9].

Crystal (1987, p. 371) in his book *The Cambridge Encyclopaedia of Languages,* as quoted by Chiswick and Miller [9], writes this regarding linguistic distance: "The structural closeness of languages to each other has often been thought to be an important factor in foreign language learning. If the foreign language (L2) is structurally similar to the original language (L1), it is claimed, learning should be easier than in cases where the L2 is very different."

B. Prosody

The goal in learning a language is for one to be able to communicate with native speakers of the target language. Classroom learning offers this communicative approach through dialogue with the teacher and other learners. This concept then, calls for TTS technology to offer the same ideology if it is to be considered for language learning. We believe that this can be achieved, i.e. TTS technology can act as a dialogue system. As it was done in the AT&T Labs in 2002 [2], TTS can be used for conversational purposes without the need to integrate a dialogue system (which would be a different matter of study and investigation). This issue then leads us to another aspect of interest: prosody.

Prosody is composed of supra-segmental features such as intonation, stress, and pitch. In order to fully meet the requirements of text-to-speech technology application in language learning, Handley [10] enforces that more attention needs to be paid to accuracy and naturalness, in particular at the prosodic level, and expressiveness.

Correct usage of supra-segmental features such as intonation and stress has been shown to improve the syntactic and semantic intelligibility of spoken language [(Crystal, 1981) as quoted by [11]]. In spoken conversation, intonation and stress information do not only help listeners to locate phrase boundaries and word emphasis, but also to identify the pragmatic thrust of the utterance (e.g. interrogative vs. declarative) [11]. One of the main acoustical correlates of stress and intonation is fundamental frequency (F0); other acoustical characteristics include loudness, duration, and tempo.

Improvement on prosody can sometimes affect other TTS features, i.e. advancement in prosody can happen at a cost of another TTS feature. For instance, Syrdal and Kim [2] pointed out that even though unit selection synthesis produces a more natural sounding speech output than its counterparts, its prosodic feature cannot be controlled. The improved naturalness provided by unit selection is achieved at the cost of the more precise prosodic control as provided by earlier, more robotic sounding synthesisers [1]. Also, Filickaya [12] clarifies that although TTS speech is human-sounding, "there is always a difference in terms of intonation and stress".

The researchers at AT&T [2] then decided to focus on the acoustic measures of prosody in a larger speech corpus from one speaker and relate this to speech acts. Speech acts are emotions or expressions shown/uttered in relation to the text being synthesised. According to [2], in dialogue systems, it would be a simple matter to convey the intended speech act to a TTS system designed to use that information at various levels in synthesising speech. Other alternatives to providing speech act information to TTS include an analysis of input text to predict the most likely speech act intended or manual text markup.

Syrdal and Kim [2] reached a conclusion that including speech act information along with input text would improve the capability of a TTS front end to assign more appropriate prosody. Our intention, as a result, is to explore the same plan and hope to obtain the same, if not better speech output.

IV. Evaluation

Evaluation of speech technology systems in language learning is still a much discussed research topic. Some learners have not embraced the technology fully and they feel as if the technology is being enforced on them. Technology developers on the other hand seem not to have figured out the proper and efficient method of evaluation.

One argument by speech technology researchers and linguists is based on whether the evaluation is on the performance of the technology itself, or the performance (competence) of learners. Since the technology is developed to enhance the performance of learners, one would base their evaluation on how learners perform, and if indeed the technology has made any significant change in their performance. At the same time, technology developers' interest will be on the efficient performance of the system.

Researchers in TTS technology have come up with various methods, a few of whom are [13] and [14]. Schroeter [13] believes that evaluation depends on which part of a TTS system impacts which criterion. For example, intelligibility and/or naturalness require conducting subjective listening tests. For overall quality evaluation, the International Telecommunication Union (ITU) recommends a specific method that [13] "is also suitable for testing naturalness". Tests such as these, Mean Opinion Score (MOS) scale for instance, have a 5 point rating scale for characteristics such as *overall impression*, *listening effort*, *comprehension*, *intelligibility*, etc.

Van Hooijdonk [14], on the other hand, performs an online evaluation, which is quite different to the 'usual' offline evaluation methods. His study was based on an eye tracking experiment to investigate the processing of diphone synthesis, unit selection synthesis, and human speech taking segmental and supra-segmental speech quality into account. His belief is that online research methods, like eye tracking, give direct insight into how listeners process speech incrementally. This also provides a better indication of segmental intelligibility of synthesised speech.

The linguist Noam Chomsky has argued that [15] "it is the business of theoretical linguistics to study and model underlying language competence, rather than the performance data of actual utterances which people have produced (Chomsky, 1965)." By competence, Chomsky is referring to the abstract and hidden representation of language knowledge held inside people's heads, with its potential to create and understand original utterances in a given language [15].

However, there are linguists who do not understand how competence can be studied. They believe that, in principle, the infinite creativity of the underlying system can never adequately be reflected in a finite data sample.

All in all, researchers need to take into consideration the fact that effective evaluation metrics can pinpoint weak or missing technologies. A lab prototype is different from a commercial system [16]. Effective data collection and performance evaluation methods play an important role in speech technology systems.

To cut the long story short, Handley suggests that evaluation process should be based on these guidelines [10]:

a) Establish the evaluation requirements (establish purpose of evaluation, identify types of products to be evaluated, specify the quality model);
b) Specify the evaluation (select metrics, establish rating levels for metrics, establish criteria for assessment);
c) Design the evaluation;
d) Execute the evaluation.

V. Other Considerations

As developers of technology which will end up in the hands of non-technical users, we need to take the idea of their 'technological illiteracy' into consideration. We may develop an optimal system which can perform brilliantly, but the same system should perform likewise from the users' perspective (based on the knowledge and skills they have). Lai and Kritsonis mentioned that research findings indicate that the use of computers has a "positive effect on the achievement levels of L2 learners, but it still has its limitations and weaknesses, such as financial, isolated, and knowledge required" [17]. For this reason, it is necessary that teachers and learners should have basic technology knowledge before they apply computer technology to assist with L2 teaching and learning. No student can utilize a computer if they lack training in the use of computer technology.

TTS technology has improved a lot and it is ready to be deployed in language learning provided its limitations are taken into consideration [12]. Basic understanding of TTS technology and its advantages and limitations in a given application is essential to end-users. If traditional teachers intend to expose students to natural language audio input and 'comprehensible input' as much as possible, TTS technology can provide a valuable way of doing it.

VI. Concluding Remarks

The focus of this paper has been an investigation into how text-to-speech technology can be optimised for second language learning. In our opinion, the best and first thing to do is to develop a TTS system with as high quality speech output as possible. From literature read so far, a lot remains to be done as the optimal mark has not been reached as yet.

Since the idea is to optimise TTS technology for second language learning, the authors put the focal point on one major factor – prosody, but in an interactive dialogue notion. The goal is to get the best prosodic speech output, while at the same time maintaining the interactive/dialogue feature of the TTS system. (It must be noted that improved prosody leads to better sounding speech output - naturalness.) The paper further proposes the use of language distance measure, which will hopefully be an essential new feature to add to our TTS system. With vigorous research and study of the proposed techniques, we believe that text-to-speech technology can be a more efficient teaching and learning tool for second language learning, much better than it presently is.

For future work, we plan to investigate coupling of language generation and TTS technology. From the point of view of [18], this integration can potentially produce higher quality

output speech than could be achieved with a decoupled system, since it permits finer control of prosody.

VII. References

[1] Speech Synthesis – Wikipedia
Available:
http://en.wikipedia.org/wiki/Speech_synthesis

[2] A. K. Syrdal, Y-J. Kim, "Dialogue Speech Acts and Prosody: Consideration for TTS", *In Proceedings of Speech Prosody 2008*, pp. 661-665, Brazil, 2008.

[3] "Improving TTS Output by Controlling Input Text", White Paper by Nuance Communications, Inc., 2006.
Available:
ftp://ftp.scansoft.com/pub/UnivIdaho/wp_RS_TTS_ControlInput.pdf

[4] Y. Kim, A. Syrdal & A. Conkie, "Pronunciation Lexicon Adaptation for TTS Voice Building", *INTERSPEECH 2004*, ICSLP, pp. 2569-2572, Korea, 2004.

[5] A. Conkie, A. Syrdal, Y-J. Kim, M. Beutnagel, "Improving Preselection in Unit Selection Synthesis", *INTERSPEECH 2008*, Australia, September 2008.

[6] A. Black, P. Taylor, "Automatically Clustering Similar Units for Unit Selection in Speech Synthesis", *In Proceedings of EUROSPEECH*, Vol. 2, pp.. 601-604, 1997.

[7] V. Karaikos, S. King, R. Clark & C. Mayo, "The Blizzard Challenge 2008", The Blizzard Challenge Workshop 2008.
Available: http://festvox.org/blizzard/blizzard2008.html

[8] N. Zulu, "Measuring Language Distance", Poster, 2007.
Available:
www.meraka.org.za/nhn/Members/meraka/student-poster-day-20-june-2007/nicholas_zulu.pdf/download

[9] B. R. Chiswick & P. W. Miller, "Linguistic Distance: A Quantitative Measure of the Distance between English and Other Languages", Discussion Paper Series, IZA DP No. 1246, 2004.

[10] Z. Handley, "Evaluating Text-to-Speech Synthesis for use in Computer Assisted Language Learning", Presentation Slides, LSRI, University of Nottingham, January 2008.
Available:
www.lsri.nottingham.ac.uk/zh/Presentations/LSRIJan08.ppt

[11] F. Ehsani & E. Knodt, *Speech Technology in Computer Assisted Language Learning: Strategies and Limitations of a new CALL Paradigm*, Language Learning Technology, Vol. 2, No. 1, pp. 54-73, 1998.

[12] F. Kilickaya, *Text-to-Speech Technology: What does it offer to foreign language learners*", ISSN 1442-438X, CALL-EJ Online, Vol. 7, No. 2, 2006.
Available:
http://www.tell.is.ritsumei.ac.jp/callejonline/journal/7-2/Kilickaya.html

[13] J. Schroeter, "Text-to-Speech (TTS) Synthesis", AT&T Laboratories.

[14] C. Van Hooijdonk, et al, "Using Eye Movements for Online Evaluation of Speech Synthesis", *In Proceedings of INTERSPEECH*, pp. 1346-1349, 2007.

[15] R. Mitchell & F. Myles, "Second Language Learning: Concepts and Issues", *A chapter in the book entitled* "English Language Teaching in its Social Context", Edited by C. N. Candling & N. Mercer, published by Routledge, 11 New Fetter Lane, London EC4P 4EE, 2001.

[16] J.G. Joseph, J. Polifroni, S. Seneff, V. Zue, "Data Collection and Performance Evaluation of Spoken Dialogue Systems: The MIT Experience", *In Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP)*, pp. 1-4, 2000.

[17] C-C. Lai, W.A. Kritsonis, *The Advantages and Disadvantages of Computer Technology in Second Language Acquisition*, Doctoral Forum – National Journal for Publishing and Mentoring Doctoral Student Research, Vol. 3, No. 1,2006.

[18] V. W. Zue, J. R. Glass, "Conversational Interfaces: Advances and Challenges", *In Proceedings of IEEE*, Vol. 88, Issue 8, pp. 1166-1180, August 2000.

Lehlohonolo Mohasi is a PhD student in Electrical Engineering at the University of Cape Town (UCT). She holds Bachelors and Masters degrees in Electrical Engineering both from UCT. Her research area is in Speech Technology. She is a member of IEEE UCT Student Branch and Women in Engineering (WIE).