

Clearer Text-to-Speech System Phonetization

Lehlohonolo Mohasi, Daniel Mashao
Department of Electrical Engineering
University of Cape Town
Rondebosch
7701

lmohasi@crg.ee.uct.ac.za, daniel@eng.uct.ac.za

Tel: +27 21 6504019

ABSTRACT

Speech is a natural form of communication for humans, but this is not so for machines. Two of the most researched categories of speech technology are speech recognition and speech synthesis. Speech recognition is the ability of a machine to convert speech into text, while speech synthesis does a reverse process. The main focus on these two is to try and get machines to be able to acquire speech (i.e. to be able to hear and give feedback) like humans. With work that has been done so far, it is possible through the use of both these technologies though not perfect. Some of the problems involved in this are intelligibility, pronunciation, naturalness, etc. We look into pronunciation which is made possible by phonetization. Phonetization of text is important for speech synthesis and recognition and for other natural language processing applications. Phonetization is conversion of letters/graphemes into phonetic sound through a set of rules and algorithms. These rules are usually language dependent, i.e. each language has its own unique set of letter-to-sound (LTS) rules. We hope to apply these rules to our Sesotho text-to-speech (TTS) system, and eventually to the Sesotho speech recognition system. Sesotho is one of the eleven official languages in South Africa.

1. INTRODUCTION

Text-to-speech systems are becoming more common. An example of a TTS system is the Telkom prepaid phone. If a user is registered by ethnic group, the user can request introduction about their balance and they will be told in their own language what their balance is. A prerecorded TTS has the advantage that it can speak clearly but it is not flexible. In designing flexible systems, pronunciation becomes important for understandability.

One approach to the transcription of written text into sounds (phonetization) is to use a set of well-defined language-dependent rules, which are in most situations augmented by a dictionary of exceptional words that constitute their own rules [2]. The transcription process starts by pre-processing the text into lexical items to which the rules are applicable. The rules can be sub-divided into phonemic and phonetic rules. Phonemic rules operate on the graphemes to convert them into phonemes. Phonetic rules operate onto the phonemes and convert them into phones or actual sounds. Converting from written text into actual sounds and developing a comprehensive set of rules for any language is marked by several problems that have

their origins in the relative lack of correspondence between the spelling of the lexical items and their sound contents.

Phonetization of text is important for both speech synthesis and speech recognition. In speech synthesis, the rules are used to derive the correspondence between the orthography and the sounds (phonemes and phones). The sounds can then be used alone or converted into syllables, which are further sub-divided into clusters to be used for the language synthesis. In speech recognition, letter-to-sound rules are used as a way of generating pronunciation variants to enhance the quality of the recognizer and generating pronunciations for new add-on words, which are not in the original vocabulary of the speech recognition system.

Part of the educational process for a child is learning to read, and this process is applied to a machine as a form of phonetization. Developing letter-to-sound rules set in software is essentially teaching the computer how to read or pronounce words in a language. The difficulty in developing an accurate algorithm to perform this task is directly proportional to the fit between graphemes and corresponding phonemes as well as the allophonic complexity of the language in question. Section 2 discusses different phonetization methods and the effect of pronunciation in automatic speech recognition (ASR) and text-to-speech systems (TTS). In Section 3, we discuss the development process of our project as we gain more phonetic knowledge. The paper ends with a brief conclusion in Section 4.

2. PHONETIZATION IN ASR/TTS SYSTEMS

2.1 Letter-to-sound transcription methods

There are three methods that have been used for the letter-to-sound transcription of most languages. They are: dictionary-based methods, rule-based methods, and trained data-driven methods.

2.1.1 Dictionary-based transcription

Dictionary-based letter-to-sound transcription relies on storing maximum phonological knowledge (including pronunciation of morphemes) in a lexicon. The pronunciation of input words is generated from the stored morphemes by complex morphological rules that include inflectional, derivational and compounding of morphophonemic rules, that describe how the phonetic transcription of the morphemic constituents vary when they are combined into words.

The lexical entries in the dictionary can have graphemic, phonetic, syntactic and semantic information. A comprehensive

dictionary requires huge computer memory and tedious effort during creation [4].

2.1.2 Rule-based transcription

Rule-based transcript systems use a comprehensive set of grapheme-to-phoneme rules, a dictionary of exceptions (words that constitute their own rules), and a phonetic post-processor to transcribe text into actual sounds. Since the emergence of rule-based methods, progressive elaborate efforts have been made to design sets of rules and exceptions of wide coverage. The most elaborate rule-based systems are expert knowledge-based systems because they use expert linguistic and phonetic knowledge to devise the rules. The different types of rule formalisms are related to the following aspects: differences in number of rules, the phonemic inventory, the types and formats of the rules, the direction in which the rules are parsed, the size of the exceptions' dictionary, the algorithm used to scan the exceptions' dictionary, etc. Rule-based methods are language specific and are widely used in speech synthesis [4].

2.1.3 Data-driven transcription

Data-driven transcription constitutes the following three approaches: pronunciation by analogy (PbA), statistical methods based on stochastic theory and nearest neighbor, and methods based on neural networks. The underlying idea in PbA is to determine the pronunciation of a novel word from similar parts of known words and their corresponding pronunciation. Thus the pronunciation of a novel unknown word is assembled by matching substrings of the input novel word to strings of known lexical words in the dictionary. In statistical methods such as stochastic transduction method, a training material like the grapheme-phoneme correspondence is used to generate a phoneme classification with a certain probability. Trained neural networks using multilayer perceptrons (MLP) and back-propagation for training have also been used in text transcription such as those developed by Sejnowski and Rosenberg (1987) [4]. MLP-based solutions are language independent but they do not handle grapheme clusters and syntactic features well..

2.2 Pronunciation

Current unit selection speech synthesis systems achieve highly intelligible and moderately natural speech [1]. Getting speech which sounds more natural requires paying more attention to linguistic phenomena such as phonetic detail. This will contribute to a better and more natural-sounding pronunciation of words for a particular language. There are methods available for measuring intelligibility, but they are not good in measuring naturalness. Measurement of naturalness is an area of ongoing research.

It is widely accepted [1] that adding some pronunciation variants to the lexicon of a speech recognizer can improve accuracy, but adding too many variants increases confusability to the point where accuracy goes down. Knowing which words to add variants for and which or how many variants to add requires some phonological

knowledge: adding phonologically well-motivated variants can increase accuracy. A recognizer which uses this information ought to be more accurate than one that does not.

3. FUTURE DEVELOPMENT

Given the nature of the Sesotho writing system and the regular relationship between its spelling and pronunciation [5 & 7], it seems that it is well suited to rule-based transcription since it is possible to develop a generalized set of letter-to-sound rules that cover the majority of Sesotho spelling. A dictionary of exceptional words will cover the exceptions to the generalized rules. In developing the grapheme-to-phoneme rules, it will be assumed that the words are spelled correctly.

Using the spelling literature of Sesotho, it is possible to compile a set of precise rules to transcribe Sesotho text phonemes and phones. Sesotho can be classified as a phonetic language having a regular spelling system. In this respect, Sesotho can be modeled with specific rules which are developed manually using our linguistic and phonetic expertise. If the list of exceptions is comprehensive and the letter-to-sound rules are complete and precise, we hope to get a precise transcription of Sesotho text. For our system, combining precise letter-to-sound rules with a dictionary of exceptional words should be enough to achieve high precision letter-to-sound transcription.

4. CONCLUSION

In this paper, we have shown that phonetic and phonological knowledge (through phonetization), if applied correctly, is beneficial to both speech recognition and synthesis systems. The use of such knowledge is required at high level if we are to achieve higher intelligibility and naturalness in our systems.

5. REFERENCES

- [1] S. King, (25 Sept. 2003). "Dependence and independence in automatic speech recognition and synthesis" *Journal of Phonetics*
- [2] Y. A. El-Imam, (8 August 2003). "Phonetization of Arabic: rules and algorithms", *Computer and Speech Language Journal*
- [3] M. Divay, A. J. Vitale, (1997). "Algorithms for Grapheme-Phoneme Translation for English and French: Applications for Database Searches and Speech Synthesis"
- [4] G. Tajchman, E. Foster, D. Jurafsky, (1995). "Building Multiple Pronunciation Models for Novel Words using Exploratory Computational Phonology"
- [5] R. A. Paroz, (1946). "Elements of Southern Sotho", Morija Printing Works
- [6] Sesotho Language, <http://www.wordspider.net/se/sesotho-language.html>
- [7] E. Jacottet, (1972). "A Practical Method to Learn Sesuto", Morija Sesuto Book Depot
- [8] Black A. W., Lenzo K., Pagel V., "Issues in Building General Letter to Sound Rules"

BIOGRAPHY

Lehlohonolo Mohasi is in her first year of MSc (Elec Eng) at the University of Cape Town. Her area of research is in Speech Technology. She did her undergrad at the same institution and graduated in 2004.