SOUNDS AND SYMBOLS: AN OVERVIEW OF DIFFERENT TYPES OF METHODS DEALING WITH LETTERS-TO-SOUNDS RELATIONSHIPS IN A WIDE RANGE OF LANGUAGES IN AUTOMATIC SPEECH RECOGNITION

Maria Goudi, Pascal Nocera

Laboratoire Informatique d'Avignon, Université d'Avignon et des Pays de Vaucluse, France

ABSTRACT

Mapping a language's graphemes to a sequence of symbols, which represent its corresponding phonemes, is of great importance for the recognition accuracy of an automatic speech recognition system. Phoneme-based and grapheme-based approaches are mainly employed for the creation of a pronunciation dictionary. In this paper, we present the application of these approaches on a variety of languages with different types of writing systems and various degrees of complexity between graphemes and phonemes.

Index Terms— Writing systems, automatic speech recognition, pronunciation dictionary, acoustic model, speech recognition, phoneme-based, grapheme-based, rule-based.

1. INTRODUCTION

The relationship between a language's spelling system and its corresponding sounds is a crucial aspect in the automatic speech recognition (ASR) process. In fact, one of the core components of the acoustic model of an ASR system is the pronunciation dictionary which is responsible for the matching between a word (as represented in the language model) and its pronunciation(s). An accurate mapping of the orthographic representation of a word to a sequence of symbols representing the pronunciation of that word is important to guarantee the recognition quality of an ASR system. By contrast, the acoustic model will be trained with wrong data and wrong models will be used during the decoding process for the calculation of scores.

Two main types of methods for generating a pronunciation dictionary are found in the ASR literature: the phoneme-based and the automatic grapheme-based ones. The phoneme-based techniques are built on the basis of pronunciation rules which are either manually created by linguists or statistically derived from available data sets [2]. These rules can be subsequently applied manually, automatically or semi-automatically.

The grapheme-based methods use graphemes as modeling sub-word units instead of phonemes. Phonemes are widely, but not exclusively, used as sub-word units in the rule-based approaches. Polyphones and syllables are also used as sub-word units in the creation of pronunciation dictionaries. The use of graphemes as modeling units does not derive from any rules. The letters of each word are directly used as the acoustic units to be modeled.

The phoneme-based methods often require access to expert knowledge of the target language for the creation of the pronunciation rules. These methods present the advantage of requiring a limited quantity of resources (i.e. corpus, etc.) on one hand, on the other, the resulting recognizer is often of better quality than recognizers based on grapheme-based techniques [8], [16], [17]. However, the phoneme-based approaches are not easily adaptable to other languages [8] and are time consuming and thus costly when it comes to the manual elaboration and testing of the designed rules [8], [16], [17]. The grapheme-based techniques are fully automatic, less costly and more easily adaptable to different systems.

As we mentioned above, in general, the grapheme-based methods do not perform as well as the phoneme-based ones, yet, some studies have shown that the results of grapheme-based recognizers depend on the nature of the grapheme-to-phoneme relation of the target language [2], [9]. In addition, in recent studies, graphemes are increasingly used as modeling units [17], consequently these techniques tend to improve rapidly.

This paper aims at giving an overview of the ways with which different methods in speech recognition deal with issues related to the letters-to-sounds relations in a wide range of languages. In the sections that follow, we shall explore some of the methods employed in order to deal with different characteristics of the main types of writing systems. In section 2, some important issues related to alphabets are

investigated. In section 3, abjads will be examined and, more specifically, the case of Modern Standard Arabic. Section 4 introduces some special features of syllabic alphabets, illustrated by the cases of Thai and Khmer. In section 5, semanto-phonetic scripts are explored. The last section is dedicated to some summing-up remarks and perspectives.

2. ALPHABETS

Alphabets generally represent vowels and consonants. They are composed of sets of letters, usually arranged in a fixed order [23]. Languages using alphabetic scripts are characterized by mainly three types of relationships between letters and sounds: a) a single letter or a group of letters represents a single sound, b) a letter or a group of letters corresponds to many sounds depending on the context or specific rules and c) a sound can be represented by a variety of letters or combination of letters. These relationships represent, in different languages, different degrees of ambiguity. Some alphabetic scripts are characterized of rather unambiguous grapheme-to-phoneme relationships. It's the case, for instance, of Czech for which rules of pronunciation can be easily defined. According to [12], some of its irregularities, being quite systematic, can be handled by regular rules and some words – mainly of foreign origin – could be treated by a special tool. Similarly, for Swahili, which is characterized of a rather regular relationship between spelling and pronunciation, a grapheme-to-phoneme script can automatically generate most of the words' pronunciations [6]. In order to handle English words and proper names, whose frequency is important enough in Swahili, pronunciation variants are added to the lexicon, based on the corresponding English pronunciations [6]. Additionally, speech recognition studies concerning Croatian [7], Russian [17] and Greek [14] show that the process of building letter-to-sound rules is rather easy to handle for these languages, as well. In English, on the other hand, the grapheme-to-phoneme relationships are more ambiguous: one letter can represent a variety of sounds conditioned by complex rules and many exceptions [22], thus, the creation of a pronunciation dictionary becomes an elaborate task.

As we have mentioned above, generally, grapheme-based approaches do not perform as well as rule-based approaches based on phonemes. However, as shown by a study carried out for Dutch, German, Italian and English, improvements to the grapheme-based techniques seem to be rather interesting, at least for languages of not very complex grapheme-to-phoneme relationships [8]. The experiments resulted in only a 2% increase of error rate for Dutch, German and Italian, languages with less ambiguous pronunciation rules, as compared to English, for which there was a 20% increase of the error rate [8]. In another study concerning English, context-dependent phonemes have been compared to context-dependent graphemes, used as sub-word units in two ASR systems [4]. The results show that in tasks of smaller complexity the grapheme-based system can perform as good as the phoneme-based one. However, in tasks of increased complexity, the phoneme-based system outperforms the grapheme-based system [4]. In the case of Afrikaans, which, according to [2], lies somewhere between the highly irregular English and the highly regular Flemish, it was shown that the performance of the grapheme-based approach is dependent on word categories. Apart from spelled out words, acronyms, proper names and foreign words, for which a degradation in word accuracy is observed, the grapheme-based system performs nearly as good as the grapheme-to-phoneme ASR and converges quickly to the performance of the manually controlled phoneme-based ASR as the training set size increases [2]. Finally, a study conducted for Russian, a language with rather unambiguous grapheme-to-phoneme relationships, showed that the grapheme-based recognizer system performed almost as well as the phoneme-based baseline system [17].

3. ABJADS

The abjads, a subcategory of alphabets also called consonant alphabets, have independent letters for consonants but vowels are often not marked. When they are marked, it is usually done by means of diacritics. Most of abjads, with the exception of Ugaritic, are written from right to left. Arabic, Hebrew and Syriac are three of the abjads still in use [23]. In this section, we have chosen to discuss the case of Arabic (and more specifically the Modern Standard Arabic), as it is a widely studied language in the ASR literature.

In the Modern Standard Arabic (MSA) alphabet, 25 letters represent consonants and 3 letters represent long vowels [20]. Short vowels and some other pronunciation information, like doubling of consonants and doubled case endings (vowels used at the end of the words to mark case distinction), are represented by diacritics [21]. The major problem that we encounter when creating an Arabic pronunciation dictionary is related to the lack of representation of the short vowels and other diacritics in its written form. Full vowel indication is only used for some political and religious texts as well as in textbooks for beginners of the Arabic language [21], [23]. Since, for a given written word, there are several possible vowelizations,

each one leading to a different meaning, this may lead to significant lexical ambiguity [20]. A non-diacritized dictionary word form corresponds on average to 2.9 possible diacritized forms [20]. In general, syntactic and pragmatic constraints may reduce ambiguity in many cases [10], but in Automatic Speech Recognition this remains a complicated task. Nevertheless, the relationship between graphemes and phonemes of a diacritized text is relatively transparent compared to other languages such as English or French [20]. According to studies so far conducted, it seems clear that phoneme-based models perform better when using diacritized data [20].

Considering the high cost for obtaining manually diacritized data, researchers have been investigating various procedures in order to automatically insert missing diacritics into available corpora, intended for acoustic model training [20], [21]. In [20] for example, diacritics are restored based on a combination of morphological and contextual constraints with acoustic information. [21] proposes a model which integrates a wide array of lexical, segment-based and part-of-speech tag elements. Finally, the approach developed by [11] combines probabilistic methods and simple linguistic information that result in high levels of accuracy (7.33 % WER) when compared with the previous studies. Yet, [5] argues that, in most studies, the accuracy of the automatic diacritization systems stays low, being in the range of 15–25% WER and thus, demanding manual reviewing in order to achieve higher accuracy.

In addition, the grapheme-based acoustic modeling approach has been explored for MSA ASR systems. Its major advantage is the fact that no diacritization is needed for acoustic training. Earlier studies, as in [1], showed that the diacritized phoneme-based approach performs much better (around 14% for [1]) when compared to the grapheme-based approach. However, in a more recent study proposed by [5], diacritics are modeled by using context-dependent acoustic models and Gaussian mixture model. By increasing the number of Gaussian densities or the amount of training data, [5] shows that the improvement rate in the grapheme-based approach is faster than in the phoneme-based approach. In that way, the accuracy gap between the two approaches could be neutralized [5].

4. SYLLABIC ALPHABETS

In syllabic alphabets, also called alphasyllabaries or abugidas, each grapheme represents one syllable. Syllables are constructed of consonants having an inherent vowel, which can be changed to another vowel or muted by means of diacritics. Diacritics are also used to separate vowel letters, when they occur at the beginning of a syllable or on their own [23]. Bengali, Khmer, Thai, Lao, Tamil, Lanna and many other languages of South and South East Asia use syllabic alphabets but research in the ASR literature is not yet very advanced for most of these languages, which are, for their majority, under-resourced.

Regarding the Thai language, [18] suggests that it is characterized of a relatively poor letter-to-sound relationship. In a relatively recent study, [3] compares a grapheme-based system with two phoneme-based ones. The results show that the grapheme-based approach, created with enhanced tree clustering, outperforms the phoneme-based approach, which uses an automatically generated dictionary, and performs almost as well as the phoneme-based system with the manually created dictionary [3].

Concerning the Khmer language, for which, on one hand, the language resources available in digital form are scarce, on the other, linguistic characteristics (especially acoustic and phonological) not yet well studied, [16] proposes a comparative study between a grapheme-based acoustic model and a phoneme-based model based on automatically generated grapheme-to-phoneme rules. The fact that the results show a comparable performance between the two models seems very promising regarding the grapheme-based approach since it could probably mean that a grapheme-based approach could be used when a hand-crafted pronunciation dictionary is not available [16].

5. SEMANTO-PHONETIC WRITING SYSTEMS

In semanto-phonetic writing systems (also called logophonetic, morphophonemic, logographic or logosyllabic), symbols may represent both sound and meaning. In this category of scripts different types of symbols are included : a. pictograms or pictographs, which resemble the things they represent, b. logograms, representing parts of words or whole words, c. ideograms or ideographs, which represent graphically abstract ideas, d. compound characters, including a phonetic element and a semantic element [23].

Chinese and Japanese are two semanto-phonetic writing systems which are currently in use. The Chinese script consists of ideograms and, mostly, semanto-phonetic compound characters [23]. Each single character represents one syllable but multiple characters correspond to one syllable, each one with a different meaning.

Regarding the automatic speech recognition systems, phonological scripts are often easier to handle than semanto-phonetic scripts in terms of generation of a pronunciation dictionary [15]. More specifically, in the case of Chinese,

character conversion tools are sometimes used as a first step before deriving the pronunciation from the converted strings [13], [15], [19]. Characters are often converted by means of the Pinyin transcription, which is a widely used transcription of the Chinese semanto-phonetic symbols using the Latin alphabet. However, the phoneme transcription is a complex task since approximately 13% of the Chinese characters have more than one pronunciation [13]. Furthermore, given the syllabic nature of the Chinese language, some studies have focused on the comparison of systems using as the basic acoustic unit either the phoneme or the syllable. In [13], for Mandarin Chinese, the results show that the phoneme-based system outperforms the syllable-based system. In addition, the expectation for a performance increase of the syllable-based approach is not as large as for the phoneme-based approach since the inter-syllable coarticulation is much smaller than the intra-syllable coarticulation [13].

6. CONCLUSION

In this paper we have tried to touch upon different aspects concerning various types of writing systems, in relation with the generation of a pronunciation dictionary. We mainly attempted to give an overview of the aspects just mentioned, therefore, many issues have not been elaborated in an exhaustive way. We plan to deal with these issues in depth in future work, giving special attention to the increasingly studied grapheme-based approach, which seems an interesting solution, not only for languages with a simple grapheme-to-phoneme relationship but, also for under-resourced languages with no available hand-crafted pronunciation dictionary.

7. REFERENCES

[1] M. Afify, L. Nguyen, B. Xiang, S. Abdou, and J. Makhoul, "Recent Progress in Arabic Broadcast News Transcription at BBN", *Proc. Interspeech*, Lisbon, Portugal, pp.1637-1640, 2005.

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

[3] P. Charoenpornsawat, S. Hewavitharana and T. Schultz "Thai grapheme-based speech recognition", *Proc. Human Language Technology Conference of the NAACL, Companion Volume*, New York City, pp. 17–20, June 2006.

[4] J. Dines and M. M. Doss, "A study of phoneme and grapheme based context-dependent ASR systems", *IDIAP Research Report 07-12,* Switzerland, March 2007.

[5] M. Elmhady, R. Gruhn, W. Minker, and S. Abdennadher, "Effect of Gaussian Densities and Amount of Training Data on Grapheme-Based Acoustic Modeling for Arabic", *Proc. IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE)*, Dalian, China, September 2009.

[6] H. Gelas, L. Besacier, and F. Pellegrino, "Developments of Swahili resources for an automatic speech recognition system", *Proc. SLTU, 3rd Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape-Town, South Africa, pp. 94-101, May 7-9 2012.

[7] I. Ipšić and S. Martinčić-Ipšić, "Croatian Speech Recognition", *Advances in Speech Recognition*, ed. N. Shabtai, InTech, pp. 123-140, 2010.

[8] S. Kanthak and H. Ney, "Context-dependent acoustic modeling using graphemes for Large Vocabulary Speech Recognition", *Proc. ICASSP*, Orlando, Florida, pp. 845-848, May 2002.

[9] M. Killer, S. Stüker, and T. Schultz, "Grapheme Based Speech Recognition", *Proc. Eurospeech*, Geneva, Switzerland, pp. 3141-3144, 2003.

[10] G. Lancioni, "Automatic extraction of prepositions in a corpus of Modern Standard Arabic written texts", *in Semitic Languages and Linguistics*, eds G. Lancioni and L. Bettini, Leiden-Boston: Brill, Vol. 62, pp. 195-211, 2011.

[11] R. Nelken and S. M. Shieber, "Arabic diacritization using weighted finite-state transducers", *Proc. ACL-05 Workshop on Computational Approaches to Semitic Languages*, Michigan, pp. 79–86, 2005.

[12] P. Pollak and V. Hanzl, "Tool for Czech Pronunciation Generation Combining Fixed Rules with Pronunciation Lexicon and Lexicon Management Tool", *Proc. LREC'02, Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain, 2002.

[13] J. Reichert, T. Schultz, and A. Waibel, "Mandarin Large Vocabulary speech recognition using the GlobalPhone database", *Proc. Eurospeech*, Budapest, Hungary, pp. 815-818, 1999.

[14] J. Riedler and S. Katsikas, ""My small slim Greek ASR system" or Automatic Speech Recognition of Modern Greek Broadcast News", *Proc. Eurospeech*, Xanthi, Greece, 2003.

[15] T. Schultz and A. Waibel, Laguage-independent and language-adaptive acoustic modeling for speech recognition, *Speech Communicaton vol 31*, pp. 31-51, 2001.

[16] S. Seng, S. Sam, V.-B. Le, B. Bigi, and L. Besacier, "Which units for acoustic and language modeling for Khmer Automatic Speech Recognition? ", *Proc. SLTU, 1st Workshop on Spoken Language Technologies for Under-Resourced Languages*, Hanoi, Vietnam, May 5-7 2008.

[17] S. Stücker and T. Schultz, "A grapheme-based speech recognition system for Russian", *Proc. 9th Conference Speech and Computer, SPECOM*, St. Petersburg, Russia, pp. 297-303, 2004.

[18] S. Suebvisai, P. Charoenpornsawat, A. Black, M. Woszczyna, and T. Schultz, "Thai automatic speech recognition", *Proc. ICASSP*, pp. 857-860, 2005.

[19] S.-C. Tseng, "Processing Mandarin spoken corpora", *Traitement Automatique des Langues*, *Special Issue: Spoken Corpus Processing*, vol. 45, nº 2, pp. 89-108, 2004.

[20] D. Vergyri and K. Kirchhoff, "Automatic diacritization of Arabic for Acoustic Modeling in Speech Recognition", *Proc. Workshop on Computational Approaches to Arabic Script-based Languages*, Coling, pp. 66-73, August 2004.

[21] I. Zitouni, J. S. Sorensen and R. Sarikaya, "Maximum Entropy Based Restoration of Arabic Diacritics", *Proc. Coling-ACL*, Sydney, Australia, pp. 577–584, 2006.

[22] A. Waibel, H. Soltau, T. Schultz, T. Schaaf, and F. Metze, "Multilingual speech recognition", Verbmobil: Foundations of Speech-to-Speech Translation, ed. Wolfgang Wahlster, Springer Verlag, 2000.

[23] Omniglot: the online encyclopedia of writing systems and languages, http://www.omniglot.com/