Automatic Speech Recognition with Graphemes and Phonemes in Portuguese

Paulo Sirum Ng
Genius Instituto de Tecnologia
São Paulo, Brazil
psirum@ieee.org

Ivandro Sanches
Centro Universitário da FEI
São Bernardo do Campo, Brazil
isanches@fei.edu.br

Abstract—The majority of contemporary Hidden Markov Model (HMM) speech recognizers use phonemes as the basic speech unit for acoustic modeling. This approach requires the existence of a grapheme to phoneme converter, or a pronunciation dictionary, in order to have the words represented, as accurate as possible, as a sequence of phonemes. A grapheme based speech recognition system avoids the need for a grapheme to phoneme converter. This simplifies the system as a pronunciation dictionary or a grapheme-to-phoneme converter may require human expert linguistic knowledge for their construction. This can also be convenient for embedded applications where the user has control over the definition of the commands to be recognized. In order to explore the phonemic orthography of Portuguese, this work presents a comparison of two speech recognition systems, one based in phonemes and the other based on graphemes as the unit for acoustic modeling.

Keywords-component; grapheme, phoneme, ASR, WER, comparison, Brazilian Portuguese.

I. INTRODUCTION

The fundamental unit in written language is a grapheme, which includes the alphabetic letters, numerical digits, punctuation marks, Japanese characters, and any other individual symbol used in written. The fundamental unit in spoken language is a phoneme. Every language has its own set of phonemes, usually about 20 to 60 [1]. Each word has its grapheme and phoneme representation. The grapheme sequence is the way the word is written and the phoneme sequence is the way the word is pronounced. In a phonemic orthography, a grapheme corresponds to one phoneme and in order to explore this issue for Portuguese, this work presents a comparison of two speech recognition systems, one based on phonemes and the other based on graphemes as the unit for acoustic modeling.

Two important components in a speech recognition system are the acoustics and the linguistics. The acoustic component is represented by the acoustic models, and the linguistic component can be represented by a grammar or a statistical language model (SLM). Usually, phonemes are used in the representation of the acoustic models, and graphemes are used in the definition of words of the grammar or SLM. The decodification process, responsible for providing the hypotheses of recognition results, will combine the information of these components to evaluate the best choice of result for a given spoken utterance at the input of the recognition process [2]. The acoustic models should represent the sound properties in different contexts, which are achieved through the use of a large quantity of speech data in a training procedure. The aim is to obtain acoustic units capable of realizing any word in the language.

In the Portuguese Language there is a certain level of correspondence between the graphemes (orthography) and the phonemes (acoustic). This level of correspondence varies among languages. For example, English has a smaller level of correspondence when compared to Portuguese or Spanish. Studies show that pronunciation dictionaries can be avoided at a small cost on recognition accuracy, dependent on the language. [3] concludes that for languages with a close grapheme-to-phoneme relation, the grapheme based speech recognizer performs as good as the phoneme based one. For example, in [4], the recognition word error rate (WER) increased about 2% (relative) for a certain corpus in Dutch and German languages, and increased above 18% for English. As for Russian, [5] showed that the WER of a grapheme based recognition system increased about 6% on read Russian newspaper article speech corpus when compared to the phoneme based baseline.

This work intends to determine the effect on WER for Portuguese when the acoustic units based in phonemes and graphemes are compared. Section 2 describes the phonemes and graphemes used in the experiment, section 3 presents the training and benchmarking data sets, section 4 describes the adopted methodology and presents the results.

II. PHONEMES AND GRAPHEMES FOR BRAZILIAN PORTUGUESE

For Brazilian Portuguese, the phonetic alphabet adopted in this work is represented by 57 symbols and it is shown in TABLE I. The pronunciation of each word in Brazilian Portuguese is then expressed as a sequence of these phonetic symbols. The phoneme based

pronunciation dictionary used in this work was manually verified by human experts. As examples, the word `carro` is expressed in the phonetic dictionary as /`k a x W`/, `casa` as /`k a z A`/, `leite` as /`l e_j C Y`/, etc.

TABLE I. BRAZILIAN PORTUGUESE PHONETIC ALPHABET.

Phonetic Symbols Occurrence Examples					
a **a**sa	a~ f**ã**	A as**a**	a~_j~ m**ãe**	a_j p**ai**	a~_w~ m**ão**
a_w m**au**	b **b**ola	C **t**ia	d **d**or	e **e**le	e~ r**e**mo
E f**e**bre	e~_j~ g**en**te	e_j l**ei**te	E_j id**ei**a	e_w m**eu**	E_w r**éu**
f **f**ogo	g fo**g**o	G **d**ia	i f**i**lho	i~ f**im**	i_w v**iu**
j ás**i**a	J ma**nh**ã	k **c**arro	l **l**ua	L fi**lh**o	m **m**anhã
n **n**ota	N ma**n**ga	o **o**sso	o~ c**om**	O s**ó**	o~_j~ p**õe**
o_j f**oi**	O_j d**ói**	o_w v**ou**	O_w v**ol**ta	p **p**ai	r ca**r**o
s o**ss**o	S **ch**á	t **t**u	u t**u**	u~ **um**	u~_j~ m**ui**to
u_j f**ui**	u_w m**ul**ta	v u**v**a	w so**l**	W carr**o**	x ca**rr**o
Y febr**e**	z a**s**a	Z **g**ente			

The phonetic symbols can be classified as presented in TABLE II. This classification is used during state-tying based on context-dependent units and was developed by linguist experts. There are, in total, 27 classes.

TABLE II. BRAZILIAN PORTUGUESE PHONEME SYMBOL CLASSIFICATION.

Class	List of Symbols
CONSONANTS	p b t d k g C G f v s z S Z x m n J N r l L w j
VOWELS	i e E a O o u i~ e~ a~ o~ u~ Y W A
DIPHTHONGS	i_w e_w E_w a_w o_w O_w u_w e_j E_j a_j o_j O_j u_j a~_w~ a~_j~ e~_j~ o~_j~ u~_j~
VOICED_CONSONANTS	b d g G v z Z m n J N l L r w j
VOICELESS_CONSONANTS	p t k C f s S x
STOPS_PLOSIVES	p b t d k g
AFFRICATES	C G
FRICATIVES	f v s z S Z x
NASALS	m n N J
LATERALS	l L
GLIDES_APPROXIMANTS	j w
BILABIAL_CONSONANTS	p b m w
LABIODENTAL_CONSONANTS	f v
DENTAL_ALVEOLARCONS	t d s z n r l
ALVEOPALATAL_CONSONANTS	C G S Z
PALATAL_CONSONANTS	J L j
VELAR_CONSONANTS	k g x N
FRONT_VOWELS	i e E i~ e~ Y
CENTRAL_VOWELS	a a~ A
BACK_VOWELS	u o O u~ o~ W
ORAL_VOWELS	a e E i o O u Y W A
NASAL_VOWELS	a~ e~ i~ o~ u~
REDUCED_VOWELS	Y W A
ORAL_DIPHTHONGS	i_w e_w E_w a_w o_w O_w u_w e_j E_j a_j o_j O_j u_j
NASAL_DIPHTHONGS	a~_w~ a~_j~ e~_j~ o~_j~ u~_j~
VOICED_PHONEMES	b d g G v z Z m n J N l L l r w j i_w e_w E_w a_w o_w O_w u_w e_j E_j a_j o_j O_j u_j
VOICELESS_PHONEMES	p t k C f s S x

On the other hand, in grapheme-based ASR systems the acoustic models are represented by graphemes. For Brazilian Portuguese, the list of graphemes and their classification is shown at TABLE III.

TABLE III. BRAZILIAN PORTUGUESE GRAPHEMIC SYMBOL CLASSIFICATION.

Class	List of Symbols
CONSONANTS	b c d f g h j k l m n p q r s t v x w z ç
VOWELS	a e i o u á é í ó ú â ê ô ã õ à ü y

This classification is used during state-tying based on context-dependent graphemes units. It is very simple, so that no prior phonetic knowledge was used. In total, there are 39 graphemic symbols and 2 classes.

Basically, the graphemic transcription of a word is its sequence of letters. This makes the grapheme based pronunciation dictionary a very simple pronunciation dictionary. All non-verbalized symbols such as hyphens (-) and apostrophes (') were suppressed. A small excerpt of the pronunciation dictionary is given below.

dispensou	d i s p e n s o u
dispensou-o	d i s p e n s o u o
dispersão	d i s p e r s ã o
displasias	d i s p l a s i a s

Portuguese speakers might say that "H", for example, is actually not pronounced when occurring in the beginning of a word. Or that, "S" between vowels within a word is pronounced like "Z". They might also mention

phonetic phenomena regarding group of graphemes such as “RR”, “SS”, “CH”, “LH” and “NH”, which have special pronunciation rules. But in order to keep the system simple with no pronunciation preprocessing, all these language specific phonetic phenomena were disregarded.

On the other hand, it was necessary to add graphemic transcriptions for letters of the roman alphabet close to their phonetic transcriptions (for spelled words) as pronounced in major part of the country. These transcriptions are given below.

b	be	j	jota	q	que	y	ipsilon
c	se	k	ka	r	erre	w	dabliu
d	de	l	ele	s	esse	z	ze
f	efe	m	eme	t	te		
g	ge	n	ene	v	ve		
h	aga	p	pe	x	xis		

III. TRAINING AND BENCHMARKING DATA SETS

The speech database used for acoustic model training consists of 114 adult speaker (57 men and 57 women) sessions totaling about 55 hours of Brazilian Portuguese single-channel close-talk pulse-code modulation (PCM) recordings at 16 kHz and 16 bits per sample. The utterances consist mostly of prompted phonetically balanced sentences, sentences from newspapers, lists of commands, numbers and names.

This database was originally used in [6] and was verified by linguist experts.

The benchmarking data consists of three different Command & Control tasks, a free-length digit sequence task and a free-length letter sequence task. They are listed in TABLE IV. All the audio files in the test set are single-channel close-talk PCM recordings at 16 kHz and 16 bits per sample. Each task is represented as an EBNF (Extended Backus-Naur Form) grammar.

TABLE IV. BENCHMARKING DATA SETS.

Test Set	Type	# utterances
CC01	Command & control – Home Automation	902
CC02	Command & control – Automotive	2496
CC03	Command & control – General	514
CDIG	Connected Digits	240
SPELL	Spelling	120

CC01 grammar consists of 94 home automation commands; CC02, 918 automotive commands and CC03, 303 general commands.

IV. EXPERIMENT METHODOLOGY

The experiment was divided in 2 parts: training and benchmarking of acoustic model based on phonemes; and training and benchmarking of acoustic model based on graphemes. The Hidden Markov Model Toolkit (HTK) [7] from Cambridge University was used for training both acoustic models.

They were trained according to the following steps:

1. Speech database preprocessing;
2. Context-independent unit estimation;
3. Context-independent based Viterbi alignment;
4. Context-independent to context-dependent alignment conversion;
5. Context-dependent unit estimation;
6. Context-dependent unit clustering (decision-tree based state-tying technique) and
7. Clustered context-dependent unit estimation and Gaussian mixture expansion.

The main HTK tools used during training were HCopy, HERest, HVite, HLEd and HHEd. In order to minimize differences as much as possible, both model sets have similar characteristics as follows:

- 13 Perceptual Linear Predictive coefficients and respective delta and acceleration coefficients;
- 3 emitting states per model;
- Strict left-to-right Hidden Markov Models;
- Context-dependent and Tied-state units;
- Diagonal covariance matrices;
- 12-Gaussian mixture per emitting state.

The main differences between their training procedures were the use of specific lexicon as described in Section II and the context-dependent unit clustering.

The decision-tree question set used for clustering the context-dependent phoneme models is the combination of the phoneme classification presented in TABLE II. and the actual list of phonemes. As for clustering context-dependent grapheme models, the question set is the combination of the grapheme classification TABLE III. and the list of graphemes.

At the end of the training part, the phoneme based HMM set has 4152 emitting states and a total of 49824 Gaussians while the grapheme based one, 3865 emitting states and 46380 Gaussians.

The HTK HVite tool was also used for model evaluation. The experimental recognition results (Word Correct and Word Accuracy rates) are presented in TABLE V.

TABLE V. BENCHMARKING RESULTS.

Test Set	Phoneme based ASR		Grapheme based ASR	
	WordCorr(%)	*WordAcc(%)*	*WordCorr(%)*	*WordAcc(%)*
CC01	96,52	95,93	95,81	95,55
CC02	85,44	84,50	85,62	85,55
CC03	98,88	98,76	98,99	98,99
CDIG	99,76	98,71	99,27	98,71

SPELL	91,28	80,53	71,21	65,35

Comparing the benchmarking results, there is no considerable difference in performance between the phoneme based speech recognizer and the grapheme based one when evaluated over Command & Control and Connected digit experiments. But grapheme based speech recognizer is considerably worse than the phoneme based over Spelling experiment.

Taking into account that Command & Control and Connected digit tasks, whose vocabularies are made of whole words, may represent more common speech recognition applications than Spelling task, the results show the equivalence of using grapheme or phoneme units.

V. CONCLUSION

The advantage of knowing that a language can avoid the usage of a grapheme to phoneme converter, or a pronunciation dictionary, is considerable. This is because the creation of such a converter, or a manually generated pronunciation dictionary, demands a considerable work. Also, the system as a whole can be simpler to implement and a user can define easily new accepted words (or commands). The results showed that the Portuguese language may allow, in some scenarios, the choice of using or not a grapheme to phoneme converter without impacting considerably the accuracy of the system. This conclusion may be convenient for applications where the user has control over the vocabulary and/or there is CPU/memory limitation.

ACKNOWLEDGMENTS

The authors acknowledge the significant support provided by Genius Instituto de Tecnologia and FINEP (Financiadora de Estudos e Projetos), ref. 3147/06, during the development of this work.

REFERENCES

[1] R. L. Trask and P. Stockwell, Language and linguistics: the key concepts. Routledge, 2nd edition, 2007.

[2] X. Huang, A. Acero, and H.-W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall, 2001.

[3] M. Killer, S. Stüker, and T. Schultz, Grapheme Based Speech Recognition, Eurospeech, Geneva, 2003.

[4] H. Ney, M. Bisani, S. Kanthak, and M. Pitz, CORETEX - Improving Core Speech Recognition Technology, 1999.

[5] S. Stüker and T. Schultz, "A Grapheme Based Speech Recognition System for Russian", Proceedings of the 9th Conference Speech and Computer, St. Petersburg, Russia, 2004.

[6] P. S. Ng and I. Sanches, "The Influence of Audio Compression on Speech Recognition Systems", Proceedings of the 9th Conference Speech and Computer, St. Petersburg, Russia, 2004.

[7] Hidden Markov Model Toolkit (HTK), http://htk.eng.cam.ac.uk