A Speech Recognition Back–End Algorithm for Portuguese Language with Unlimited Vocabulary

Francisco J. Fraga

Department of Telecommunications
Instituto Nacional de Telecomunicações, Santa Rita do Sapucaí - MG, Brazil
Telephone: +55 35 3471 9354 Fax: +55 35 3471 9314 E–mail: fraga@inatel.br

ABSTRACT

It is possible to implement a speech–to–text system with unlimited vocabulary by connecting two subsystems: A phoneme recognizer, which is performed by sub–syllabic segmenting the incoming speech, and a phonologic–graphemic converter. This paper presents an automatic speech recognition system with these features. The segmentation method and the phoneme recognizer are briefly described while the phonologic–graphemic converter is detailed. The algorithm that allows the transition from the phoneme level to the word level is based on rules obtained from the structure of the Portuguese language. This task is achieved without any kind of pronouncing tables, which allows the system to recognize any word that belongs to the Portuguese lexicon, without limitation on the size of the vocabulary

I. INTRODUCTION

As soon as we think about building an automatic speech–to–text system, we set the question of which phonetic units will be used. The HMM technique works better when the phonetic units are whole word models [1], but the amount of training data required necessarily limits the vocabulary size to some hundreds of words. The solution, of course, we find by using sub–word models, like syllables, semi–syllables or phones. However, when the phonetic units were not whole words, the error rate is almost always increased. This occurs because inter–word phenomena such as nasalization and co–articulation between phones are difficult to model [2].

The phonetic structure of Portuguese language spoken in Brazil shows that the vowels have a remarkable role in that language (vowels occur more often in Portuguese then in other languages). This fact suggests the use of vowel–based phonetic units like syllables, since the vowel is always the syllabic center in the Portuguese language. Furthermore, every syllable can be divided in two semi–syllables: The first one, called ascendant, is the initial part, from the beginning of the syllable to the center of the vowel. The second part is called descendant semi–syllable, which is formed by the second half of the vowel and the subsequent consonants, if they exist. The Portuguese language has no more than 700 semi-syllables. If we divide this set in groups corresponding to each vowel, we get sub-vocabularies with less then 100 words each, which is the ideal size for the use of the HMM technique.

Another aspect related with vocabulary size is the question of lexical entries. Some automatic speech recognition systems specially developed for Portuguese language and already successfully implemented[1], can recognize around 60,000 words. In such systems, the user can increase the vocabulary, including new words, one by one. Nevertheless, even if in this way a large vocabulary can be created, it never will be unlimited, although being flexible, because each new word must be singly added on.

This procedure is related with the speech recognition method used by these systems. The most used method, which presents better performance, does an association of a hidden Markov model with each phonetic unit [3]. In general, the phonetic units are phones corresponding to all phoneme realizations from a language.

For large vocabulary systems, in addition to the word models formed by phonetic units' concatenation, they used to use a statistic language model. The language model assigns probabilities to events corresponding to word strings that make sense in that language [4]. These models can reduce the perplexity[2] to some dozens, increasing significantly the recognition rates and reducing the computing time spent by the searching algorithm.

In order to know which units must be concatenated to form a word, they need to generate lexical entries for each word belonging to vocabulary. The lexical entries are phone strings that indicate to the recognition engine which phonetic units ought to be concatenated so as to obtain each vocabulary word.

In modern speech recognition systems, which allow the user to add on new words to vocabulary, the lexical entries are automatically generated from word spelling. To do this, they employ the same algorithm used by text–to–speech softwares, which can achieve lexical

[1] For example, the *IBM Via Voice*.

[2] Perplexity (PP) derives from entropy (H) by the formula PP = 2^H. The perplexity, when applied to language models, indicates the mean number of words that can follow a previously determined word.

entries (phone strings) from word spelling [5]. A way of transforming these large vocabulary systems in unlimited vocabulary systems would be developing an algorithm that makes just the opposite: Starting from a lexical entry, i. e., a phone string, and generating the word spelling.

The purpose of this work was to investigate the possibility of developing such algorithm; which would be able to convert a phone string in one or more grapheme strings, without using any kind of lexical entries, as it is usually done in large vocabulary speech recognition systems [6]. With the aim of achieving it, we discovered specific rules, which were applicable to Brazilian Portuguese language, of transforming phone or phoneme strings in grapheme strings, in order to enable using an unlimited vocabulary.

Section 2 brings a brief explanation of the phoneme recognition step, through the segmentation of the speech signal into semi-syllables. In section 3 we present the algorithm that converts phoneme strings in possible grapheme strings that will be related with words from Portuguese language. Section 4 deals with the obtained results on speech–to–text conversion and finally section 5 makes the conclusion, remarking the advantages of using this phonologic–graphemic conversion algorithm in large vocabulary speech recognition systems.

II. PHONEME RECOGNITION

A. Phonemes and semi-syllables

The approach employed in phoneme recognition was segmenting each word in sub-syllabic units, which are further converted into a phonemic sequence. By means of this previous segmentation, the continuous speech recognition can be done by the same methods utilized for isolated word recognition, without the use of searching algorithms, such as *Level Building* [7].

It is not the purpose of this paper to do a detailed description of the method employed on phoneme recognition step. The detailed description of the complete system was presented in a doctorate thesis [8]. The approach utilized in the referred system was to segment each word presented to the recognizer in semi–syllables, which were then converted into a phoneme sequence that also contain an apostrophe (‘) indicating the accent’s position within the sequence.

On the other hand, any other system which is able to recognize phonemes from Brazilian Portuguese language, can be used before the phonologic–graphemic conversion algorithm, since it uses the same characters showed on the second column of Table 1 for representing phonemes. This table presents a list of all phonemes from Portuguese language, as it is spoken in Brazil. In the first column, each phoneme is represented in IPA (International Phonetic Alphabet) characters. In the second column the same phoneme is represented as it was used in the computational implementation of the phonologic–graphemic conversion algorithm (OR – Our Representation). The third column presents a Portuguese word example, which contains that phoneme, with the corresponding letters printed in boldface.

For computational convenience, the vowels nasalization, revealed after them by the nasal archiphoneme / *N* / in the first column, were indicated before the vowels by the character ~ in the second column. Although in this case we have two phonemes (a vowel followed by the nasal archiphoneme / *N* /), in the algorithm’s implementation we considered them as a single different phoneme, represented by the character ~ plus the character representing the oral vowel.

TABLE 1: Portuguese language phonemes

Phoneme IPA	**Phoneme OR**	**Word Example**	**Phoneme IPA**	**Phoneme OR**	**Word Example**
/a/	**a**	b**a**sta	/*k*/	**k**	care**c**a
/ɛ/	**E**	d**e**la	/*l*/	**l**	gaze**l**a
/*e*/	**e**	m**e**smo	/ʎ/	**L**	mu**lh**eres
/*i*/	**i**	d**i**zem	/*m*/	**m**	nenhu**m**a
/ɔ/	**O**	f**o**rte	/*n*/	**n**	**n**ada
/*o*/	**o**	b**o**nito	/ɲ/	**N**	so**nh**e
/*u*/	**u**	g**u**ri	/*p*/	**p**	**p**otente
/a*N*/	**~a**	m**an**da	/ɾ/	**r**	ca**r**o
/*eN*/	**~e**	hom**en**s	/*r*/	**R**	ca**rr**o
/*iN*/	**~i**	**im**porta	/*s*/	**s**	bon**s**
/*oN*/	**~o**	c**on**forto	/*t*/	**t**	**t**ema
/*uN*/	**~u**	**un**s	/*v*/	**v**	jo**v**em
/*b*/	**b**	**b**ula	/ʃ/	**x**	**ch**efe
/*d*/	**d**	juran**d**o	/*z*/	**z**	pe**s**a
/*f*/	**f**	**f**enda	/*y*/	**y**	mã**e**
/*g*/	**g**	**g**onzos	/*w*/	**w**	pã**o**
/ʒ/	**j**	**g**erentes			

The advantage of using semi–syllables become clear when we observe that all descendant semi–syllables in Portuguese language ends in a vowel (followed or not by /*r*/ or /*s*/) or in diphthong (followed or not by /*s*/). In other words, there are few possible combinations for the descendant semi–syllables.

Furthermore, several vowel–consonant and consonant–vowel combinations are forbidden in

Portuguese. As an example, if we do not consider the diphthongs, some of these forbidden combinations (for syllables of type CVC) are:

1. All the nasal vowels cannot be followed by /*r*/;
2. /ɲ*iN* / , /ʎ*oN* / and /ʎ*oN* / do not exist.

With the aim of obtaining the implementation of all speech-to-text conversion steps, with an unlimited vocabulary, we had to introduce some restrictions:

1. Clear and paused word pronouncing ;
2. Elimination of diphthongs.

These restrictions were done in order to facilitate the segmentation task, which requires the identification of the central vowel from each syllable, and to reduce the total number of semi-syllables. Other reason for working with that restrictions was the limited amount of available resources for our research. But it is important to remark that the restrictions were introduced only for the segmentation step. The phonologic–graphemic conversion algorithm, which will be explained in next section, can deals with any phoneme sequence from Portuguese language.

B. Speech database and speech processing

The training database is key point for the performance of speech recognizer based on HMM, mainly if we employ continuous probability density functions. To effectively train the sub-syllabic models, hundreds utterances of each syllable were necessary [2], all of them were pronounced by a single male speaker. The utterances were analyzed using a Hamming window of 20 ms, with 50% of superposition (10 ms frame rate). For each speech frame, 12 mel-cepstral coefficients [9] were extracted, with the normalized log-energy appended. The first derivative from all the 13 vector components was calculated and appended, forming an observation vector of 26 components.

C. HMM topology

Three models were considered for each syllable: a vowel model, for the syllabic central region, an ascendant semi-syllable model and a descendant semi-syllable model. For all of them we utilize the left-right topology as illustrated in Figure 1. The two semi-syllables were modeled by a four state HMM, while the vowel was modeled by a three state HMM

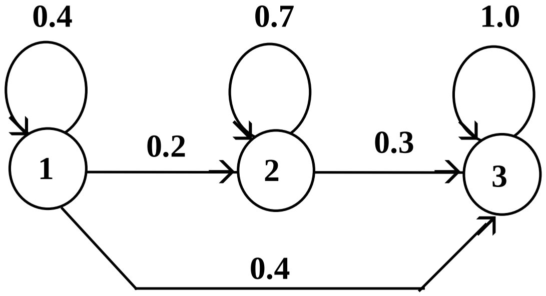

Figure 1 : Vowel HMM

III. PHONOLOGIC–GRAPHEMIC CONVERSION ALGORITHM

Before starting the phonologic–graphemic conversion, phonemes are previously analyzed and with some restrictions imposed to the recognized phoneme string, we get an initial depuration. This depuration allows the elimination of phoneme sequences that are forbidden in Portuguese. Afterwards, phonemes are classified in three types, according to their position into the incoming string: Initial, medial and final ones.

Specific rules are then considered for each phoneme and its context, i.e., the previous and subsequent phonemes within the word. Based on Linguistic and Philology knowledge concerning to Brazilian Portuguese [10], several grapheme strings are produced from a single incoming phoneme string. These grapheme strings are possible spellings of the spoken word, which are ordered by probabilistic criteria extracted from lexicon.

In the algorithm structure, we considered also the characteristic accent from some Brazilian regions. We have to deal only with those accents that produce differences in phonologic transcription, because the phoneme recognizer absorbs the ones belonging to phonetic level during the transition to phonologic level. As a pronouncing difference example on phonetic level, we have the Portuguese word “tia”: The initial phoneme / *t* / can be spoken using the plosive [*t*] or its fricative form [*tʃ*]. If we replace [*t*] by [*tʃ*] we do not change the meaning of the word “tia” nor form an other phoneme; in this context we say that the allophones [*t*] and [*tʃ*] are different realizations of the phoneme / *t* / [11].

On phonologic level, we have words that permit pronouncing variations like “mentira”, spoken carefully as / *meNt’ira* / or more often as / *miNt’ira* /; and “homem”, pronounced as / ‘ɔ*meyN* / or differently as / ‘*oNmeN* /. The algorithm was developed considering also these pronouncing variations on phonologic level. In fact, they do not inhibit the correct word spelling achievement, as we can see in the examples showed in Table 3.

In order to demonstrate the complexity involved in the algorithm development, we present in Table 2 the rules related with phoneme / *s* / when it appears in the middle of the word (the sign ! before some rules indicates logical negation). The 2I and 3I commands from Table 2 produce graphematic possibilities multiplication, even if there was a single incoming phoneme string. Indeed, the number of different graphematic possibilities generated by the algorithm varies according to the phoneme and its context. This can be observed in Table 2, on decision rules where the previous and/or subsequent phonemes are investigated.

TABLE 2: Rules for medial phoneme / s /

R1	Subsequent phoneme is vowel or semivowel
R2	Previous phoneme is oral vowel
R3	Word starts with /*e*/ or /*ine*/
R4	Subsequent phoneme is /ɛ/
R5	Subsequent phoneme is /*e*/, /*i*/, /*eN*/, /*iN*/ or /*y*/
R6	Subsequent phoneme is /*oN*/ or /*uN*/
R7	Subsequent phoneme is /ɛ/, /*e*/ or /*eN*/
R8	Subsequent phoneme is /*i*/ or /*iN*/
R9	Prev. phoneme is nasal vowel, semivowel or /r/
R10	Previous phoneme is /*b*/
R11	Previous phoneme is nasal vowel
R12	Subsequent phoneme is /ɛ/, /*e*/, /*i*/, /*eN*/, /*iN*/, /*y*/
READ MEDIAL PHONEME / *s* /	
If	!R1 → 2I (**x , s**)
If	R1R2!R3R4 → 2I (**c , ss**)
If	R1R2!R3!R4R5 → 3I (**c , sc , ss**)
If	R1R2!R3!R4!R5!R6 → 3I (**ç , sç , ss**)
If	R1R2!R3!R4!R5R6 → I (**ss**)
If	R1R2R3R7 → 2I (**xc , ss**)
If	R1R2R3!R7R8 → 3I (**c , sc , ss**)
If	R1R2R3!R7!R8!R6 → 3I (**ç , sç , ss**)
If	R1R2R3!R7!R8R6 → I (**ss**)
If	R1!R2R9!R12 → 2I (**ç , s**)
If	R1!R2!R9R10 → 3I (**c , sc , s**)
If	R1!R2R9R12R11 → 3I (**c , sc , s**)
If	R1!R2!R9!R10!R12 → I (**s**)
If	R1!R2!R9!R10R12 → I (**c**)

Table 3 illustrates several phoneme sequences submitted to the phonologic–graphemic conversion algorithm, and its respectives outgoing grapheme sequences. It is interesting to observe that, in all cases presented in Table 3, every spelled word (graphematic possibility), even the incorrect ones, when pronounced, converges to the same incoming phoneme string. This fact explains partially why it is so easy to make mistakes on word spelling: Actually, some words could be written in many different ways. In other words, the correct spelling is a convention, determined by philological and historical reasons.

TABLE 3: Phonologic–graphemic conversion examples

Input : **/`Om~ey/** Output : omem **homem** ómen hómen	Input : **/m~it`ira/** Output : mintira **mentira** mintera mentera	Input : **/as`Esu/** Output : **acesso** hacesso assesso hassesso aceço haceço asseço hasseço acesço hacesço assesço hassesço acessu hacessu assessu hassessu aceçu haceçu asseçu hasseçu acesçu hacesçu assesçu hassesçu
Input : **/`~om~e/** Output : omem **homem** ômen hômen	Input : **/awz`~eti/** Output : ausênti hausênti alsênti halsênti auzênti hauzênti alzênti halzênti **ausente** hausente alsente halsente auzente hauzente alzente halzente	
Input : **/ses`~aw/** Output : **sessão** **cessão** **seção** ceção sesção cesção		

We can observe just the opposite in some situations, i.e., different incoming sequences generate the same outgoing word, for example, the word "**homem**", which is presented in Table 3 as a same output from two different phoneme strings. Already in Table 3, the correct word spelling was signaled in boldface by software[3] (orthographic correction). This orthographic correction constitutes the last step of the complete speech–to–text system, for Brazilian Portuguese language, with unlimited vocabulary [8].

The algorithm probabilistically orders the outputs showed in Table 3; the first ones are those more often related with the corresponding input. To obtain these hierarchy and the phoneme–letter conversion rules, an extensive research and classification work was done, based on the whole Portuguese lexicon [12].

For the phonologic sequence **`/ses`~aw/`** presented in Table 3, more than one correct word was signaled. In this and in other few cases, the final decision only could be done through a semantic analysis of the words and its contexts.

IV. RESULTS

The speech database used for testing the recognition rate of the complete speech-to-text system is composed by 200 phonetically balanced phrases, slowly spoken by a male speaker. In these phrases there are 1729 words and 6988 phonemes, 3496 of them are vowels and 3492 are consonants. With the purpose of improving the recognition rate, for each spoken word, the phoneme recognition step delivers more than one possible phoneme sequence to the phonologic-graphemic converser.

First of all we show (Table 4) the results for the ***n*** first possible spelling words (only the right spelled ones) generated by the system, for each incoming spoken word from each phrase.

TABLE 4: Recognition rates for the first right spelling words

n	***Phonemes Accuracy***	***Words Accuracy***	***Phonemes Insertion***	***Phonemes Exclusion***
1	**95.9%**	**87.0%**	**0.72%**	**1.07%**
2	98.0%	92.8%	0.66%	0.72%
3	98.6%	94.4%	0.63%	0.69%
6	99.0%	96.5%	0.55%	0.61%

Next we present (Table 5) the recognition rates without taking care of the right spelling, but taking the ***n*** former graphematic possibilities generated by the algorithm.

TABLE 5: Recognition rates for the former outgoing words

n	***Phonemes Accuracy***	***Words Accuracy***	***Phonemes Insertion***	***Phonemes Exclusion***
1	**87.9%**	**60.6%**	**1.33%**	**1.08%**
2	90.7%	67.8%	1.33%	1.05%
3	91.8%	71.0%	1.33%	1.02%
6	95.2%	81.4%	1.30%	0.93%

We can note in the first line of each table that the phonemes accuracy is higher than the words accuracy. But it grows significantly when we take more possibilities, specially using orthographic correction (Table 4). In this way, it is important to note a great difference among our system and those that use lexical entries and searching algorithms (usually with a language model too), where the phonemes accuracy is always lower than the words accuracy [13]. This is a self-characteristic of large (but limited) vocabulary systems, where the words or phrases are formed by phonetic units concatenation and they seek out the correct phrase by means of searching algorithms and language models.

Just the opposite, in our system, with unlimited vocabulary, the words accuracy is lower than the phonemes accuracy. It happens because it is not possible to restrict the looking for the right words unless using the orthographic correction, just successfully implemented, and the semantic and syntactic analysis, suggested for future works. Furthermore, we want to call attention for the last line of Table 4: The remarkable improving on recognition rates with the increasing of the considered possibilities ***n*** shows that, if we could choose always the best among them, we would reach high word recognition rates. Therefore, we believe that the high rates (phonemes accuracy of **99.0%** and word accuracy of **96.5%**) obtained when we take the best among the first six possibilities, could be reached by the same system, if we submit the final text to a semantic and syntactic analysis.

V. CONCLUSION

In this paper we described a speech–to–text system, with previous speech segmentation into semi–syllables before proceeding to phoneme recognition. Afterwards, we showed the development of an algorithm that converts a phoneme sequence in one or more grapheme sequences. It is entirely based on rules extracted from Portuguese language structure, which allows the transition from phonologic level to word level, without using any kind of lexical entries. With a previous phoneme recognition step and a subsequent orthographic correction step, we reach an acceptable

[3] “ProVerb” version 2.0 from Brazilian company “PC Software”

word accuracy rate, bearing in mind that the vocabulary is unlimited.

For future works, we consider the possibility of increasing the recognition rates if the outgoing text would be submitted to a post-processing step, when a semantic and syntactic analysis would be done. This post-processing could be based on natural language processing [14], in a work similar to the one realized by Mudler [15] for German language.

VI. REFERENCES

[1] L. R. Rabiner, B. H. Huang; "An Introduction to Hidden Markov Models", *IEEE ASSP Magazine*, january 1986.

[2] L. R. Rabiner, B–H. Juang; *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[3] X. D. Huang, Y. Ariki, M. A. Jack; *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, 1990.

[4] CHIEN, L. F. ; CHEN, K-J. ; LEE, L-S. "A Best-First Language Processing Model Integrating the Unification Grammar and Markov Language Model for Speech Recognition Applications". *IEEE Transactions on Speech and Audio Processing*, vol. 1, nº 2, pp 221-239, April 1993.

[5] VAN COILE, B. "On the Development of Pronunciation Rules for Text-to-Speech Synthesis". *Proceedings of Eurospeech Conference*, Berlin, Sep 1993, pages 1455-1458

[6] ZHAO, Y. "A Speaker-Independent Continuous Speech Recognition System Using Continuous Mixture Gaussian Density HMM of Phoneme-Sized Units". *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 1, nº 3, pp 345-361, July 1993.

[7] LEE, C–H., RABINER, L. "A Frame–Synchronous Network Search algorithm for Connected Word Recognition", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 37, n.º 11, pp 1649–1658, November 1989.

[8] FRAGA, F. J.; SAOTOME, O. *Conversão Fala-Texto em Português do Brasil Integrando Segmentação Sub-Silábica e Vocabulário Ilimitado*. Doctorate thesis, ITA, 1998.

[9] DAVI, S. ; MERMELSTEIN, P. "Comparision of Parametric Representations for Monosyllabic Word Recognition". *IEEE Trans. ASSP*, vol. 28, pp 357-366, 1980

[10] SILVA,M.C.; KOCH, I.G. *Lingüística Aplicada ao Português: Morfologia*, Cortez, 1983.

[11] LYONS, J. *Introduction to Theoretical Linguistics.* Cambridge University Press, Cambridge, 1968.

[12] FERREIRA, AURÉLIO B. H. *Novo Dicionário da Língua Portuguesa*, Nova Fronteira, 1975.

[13] MORAIS, E. S.; VIOLARO, F. "Sistema Híbrido ANN-HMM para Reconhecimento de Fala Contínua". *Anais do XV Simpósio Brasileiro de Telecomunicações*, pp 117-120, Setembro de 1997.

[14] PEREIRA, F.C.N.; GROSZ, B. J. *Natural Language Processing*, Elsevier, 1993.

[15] MUDLER, J. "A System for Improving the Recognition of Fluently Spoken German Speech". *Proceedings of IJCAI*, pp 633-635, 1983.