INTRODUCING A NEW PHONETIC MODEL FOR CONTINUOUS SPEECH RECOGNITION SYSTEMS

Rubem Dutra Ribeiro Fagundes,[1], *Juarez Sagebin Corrêa*[2], *Pierre Dumouchel*[3]
rubemdrf@attglobal.net *jsagebin@crt.net.br* *pdumouch@crim.ca*

Laboratório de Processamento de Sinais e Sistemas
Escola Politécnica – Universidade de São Paulo - Brasil

Speech Recognition and Understanding Group
CRIM - Centre de Recherche Informatique de Montreal - Canada

ABSTRACT

The main goal of this work is to describe a new model for a large vocabulary continuous speech recognition system using a phonetic-phonological approach. This work proposes a statistical phonetic structure, applied at the phonetic-phonological level, to improve the speech recognition performance in systems with phonetic-phonological modeling. It is showed that the general likelihood scores are increased, indicating better recognition performances. This is due to the fact that the statistical phonetic structure will lead to enhance some frequent phonetic combinations from the language itself. Such structure should be considered as an additional knowledge base, containing information about the real language phonetic structure. Also this new phonetic-phonological approach should be strongly recommended to use in spontaneous speech recognition systems.

1. INTRODUCTION

Many researchers consider the speech communication process as a distinct evolutive differential, putting man apart from others species on Earth. Towards the creation of a real speech interaction between man and machine, rebuilding this process is needed, recreating, in an artificial way, every step involved in the speech communication process.

This work was developed in the GREP research group - Groupe de Reconaissance Automatique de la Parole (Speech Recognition and Understanding Goup) at CRIM - Centre de Recherche Informatique de Montreal, Canada, and concluded in the LPS - Laboratório de Processamento de Sinais (Signal Processing Laboratory) at POLI/USP – Escola Politécnica da Universidade de São Paulo, Brazil.

1.1 – Machine Speech Interfacing System

A man-machine speech interfacing system is an automatic speech recognition/synthesis system that artificially recreates the speech communication process steps. The following topics will briefly discuss some parts of this system. For a detailed description, please refer to: [1][2].

1.1.1 Acoustic level

The acoustic processor is directly responsible for the acoustic pattern recognition of the speech signal, obtained from the signal acquisition step. In this level, the speech signal will be converted into a sequence of acoustic symbols, usually called phonemes, generating a phonetic transcription of the input speech signal. This task uses the Hidden Markov Models (HMM) well-known approach, so training one HMM structure for each speech pattern to be recognized by the acoustic pattern matching process. An extensive description of HMM is out of the scope of this paper. For a complete presentation of this subject, please refer to: [3], [4], [5], [6].

1.1.2 Phonetic - Phonological Level

This level is built upon the words found in the sentence using the sequence of phonemes created in the acoustic level. Here, the knowledge base used to represent the phonetic-phonological information, is the phonetic graph.

In this kind of representation, the phonetic transcriptions are branches in a tree graph, where each phoneme is noted by ***(in, p, fn)***, being ***in***, ***p***, ***fn*** indexes of initial node, phoneme symbol and final node respectively [7] [8] [9]. At the end of each branch, in the final node, there is a vocabulary word index; therefore the whole path from the root node to every final node represents the complete phonetic transcription of that word.

[1] ***Professor at DEE/FENG/PUCRS – Electrical Engineering Dept. Pontifical Catholic University of Rio Grande do Sul, Brazil***
[2] ***Chief Learning Officer at CRT Brasil Telecom ,Porto Alegre, Brazil.***
[3] ***Leader Researcher at CRIM – Centre de Recherche Informatique de Montreal, Canada***

This approach uses a specific graph called "*Lexical Tree*" first proposed by Herman Ney [7]. This graph is conceptually equal to a phonetic graph, emphasizing those groups of similar phonemes found in the vocabulary words. An example of this Lexical tree can be seen in figure (f.1):

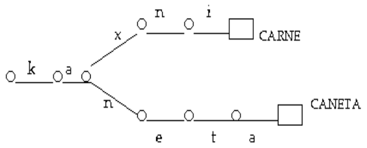

Figure f.1: Example Lexical Tree (in Portuguese)

Figure (f.1) shows that, there is one common branch for a given sequence of common phonemes, therefore relating vocabulary words with the same initial (one or two) phonemes, saving storage space and searching time.

1.1.3 Syntax level

This level uses a knowledge base with grammatical rules representing the general structure of the sentences that can be extracted from speech utterance. In large vocabulary continuous speech recognition systems, the bigram or trigram language modeling is usually applied, with formulation as seen below [10], [11], [12]:

bigrams:

$$P(w_n \mid w_1 w_2 w_3 ... w_{n-1}) \cong P(w_n \mid w_{n-1}) \qquad \textbf{(1)}$$

trigrams:

$$P(w_n \mid w_1 w_2 w_3 ... w_{n-1}) \cong P(w_n \mid w_{n-2} w_{n-1}) \qquad \textbf{(2)}$$

2. TOP-DOWN AND BOTTOM-UP APPROACHES

In the computing processing point of view, the whole task performed by a speech recognition system can be described as a search algorithm running over levels of knowledge bases, relating phonemes, words and valid hypothesized sentences generated as results from computing process. Nevertheless, this kind of computing process can be performed in two different approaches: Top-Down and Bottom-Up.

The Top-Down approach will generate, at first, many hypothetical sentences, oriented by language modeling. These hypotheses will guide the next steps at the phonetic-phonological level and acoustic level, matching words and phoneme sequences with the best likelihood score computed by the search algorithm.

On the other way, in the Bottom-Up approach, the acoustic level will perform phoneme detection, generating a phoneme sequence that will be treated at the phonetic-phonological level and at the syntax level in order to build a valid speech sentence.

2.1 Block-Diagram of a Phonetic - Phonological Speech Recognition System

The following block diagram represents a speech recognition system for a large vocabulary, where words are represented by their phonetic transcriptions:

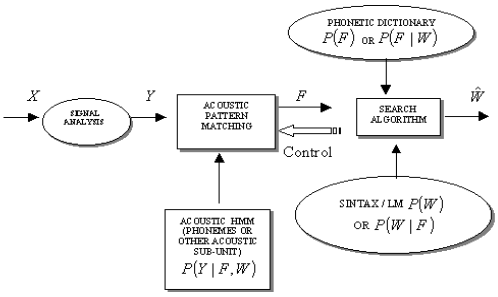
PHONETIC DICTIONARY
P(F) OR P(F | W)
X
SIGNAL ANALYSIS
Y
ACOUSTIC PATTERN MATCHING
F
Control
SEARCH ALGORITHM
Ŵ
ACOUSTIC HMM (PHONEMES OR OTHER ACOUSTIC SUB-UNIT) P(Y | F,W)
SINTAX / LM P(W) OR P(W | F)

Figure f.2 : Phonetic-Phonological Speech Recognition System

This block diagram describes, in detail, what really happens during speech recognition, computing on Top-Down and Bottom-Up approaches.

On a Top-Down approach, a set of valid *W* sentences will be generated before phoneme sequence *F* because those hypothetical *W* sentences, created by syntax processor using a given language modelling $P(W)$, will guide the acoustic pattern matching to generate the best phoneme sequence *F*. In this case, the recognition process can be described with the following expression:

$$\hat{W} = \arg\max_{W} P(Y \mid F, W) P(F \mid W) P(W) \qquad \textbf{(3.I)}$$

where $P(Y \mid F, W)$ is related to the acoustic pattern matching step, and $P(F \mid W)$ is a phonetic-phonological knowledge base, usually a phonetic dictionary, connecting the phoneme sequence *F* to the word sequence *W*.

On a Bottom-Up approach, the phoneme sequence *F* will be created before the word sequence *W* is, because the phoneme sequence will guide, at first, those others steps in a phonetic-phonological level and syntax level through the final word sequence *W*.

In this case, the recognition step is formulated by:

$$\hat{W} = \arg\max_{W} P(Y \mid F, W) P(W \mid F) P(F) \qquad \textbf{(3.II)}$$

where $P(W \mid F)$ is related to the syntax and phonetic-phonological levels, usually a phonetic graph, and $P(F)$ is some sort of phonological rule, or again, either a phonetic graph enhanced by a phonetic-phonological knowledge provided by the vocabulary, or a previous knowledge from the language used in the speech. The *lexical tree* is an example, because those groups of initial similar phoneme sequences are in fact a different way to represent this specific knowledge about the vocabulary words used in the language.

Furthermore, it should be noted that the terms $P(F|W)$ or $P(F)$ are freely exposed to point out the fact that by handling these terms it is possible to figure out new ways to improve the whole speech recognition performance. The current knowledge base presented until now (the phonetic dictionary, phonetic graphs, phonological rules) is actually a static way to represent the vocabulary, just relating words to their phonetic transcriptions.

2.2 Statistical Phonetic Modeling (SPM)

This phonetic modeling is a new way to enhance $P(F|W)$ or $P(F)$ using the knowledge provided by the language itself. This knowledge does not come from phonological rules, but is dynamically extracted from the language used by speakers in their ordinary communication, applying the same method to generate a language model. In this case, those terms will be statistically formulated as phonetic bigrams or trigrams, as follows:

Phonetic bigram:

$$P(f_n \mid f_1 f_2 f_3 \cdots f_{n-1}) \cong P(f_n \mid f_{n-1}) \qquad \textbf{(4)}$$

Phonetic trigram:

$$P(f_n \mid f_1 f_2 f_3 \cdots f_{n-1}) \cong P(f_n \mid f_{n-2} f_{n-1}) \qquad \textbf{(5)}$$

This new statistical phonetic modelling will enhance the whole continuous speech recognition performance by adjusting the likelihood score computed during the recognition step. Therefore, these conditional probabilities bigrams or trigrams will add the knowledge about the idiom, emphasizing frequently used pairs or triplets of phoneme sequences and de-emphasizing other ones not so frequent, or even wrong phoneme sequences.

3. METHODOLOGY

The basic theory and implementation technique for the acoustic pattern recognition is the well-known Hidden Markov Model (HMM). This approach will lead to new implementation ways, using new ideas at the phonetic-phonological level

A portion of the English speech corpus SWITCHBOARD (SWB), distributed by LDC (Linguistic Data Consortium), was used with a phonetic dictionary of 18000 words and a language modelling from CLSP[1]. It should be noted that (SWB) is a spontaneous speech database, and therefore, it is very hard to predict an adequate performance since spontaneous language is still a challenging problem in speech processing.

For algorithms and training implementations, the HTK toolkit version 2.1 (from *Entropics Inc.)* was used in the creation and execution of the whole recognition system.

3.1 The SPM Implementation

From the previous analysis of the expressions (3.I) and (3.II) in 2.1 we can associate the terms $P(W|F)$ and $P(W)$ as $P(\omega)$ [2], the terms $P(F)$ and $P(F|W)$ as $P(f)$ [3] With this new arrangements the expression (3.I) and (3.II) becomes a more generic expression:

$$\hat{W} = \arg\max_{W} P(Y|F,W)P(\omega)P(f) \qquad \textbf{(6)}$$

or using logarithmic values:

$$\hat{W} = \arg\max_{W} \{\log[P(Y|F,W)] + \log[P(\omega)] + \log[P(f)]\} \qquad \textbf{(7)}$$

3.1.1 The α contribution parameter

To check how important the influence of $P(f)$ is, in the global recognition score, it will be necessary to re-scale the terms of (7), since that, from computing process, each score is calculated from SPM and it will be as large as any other term in that expression. The parameter α (with a range from 0.01 to 0.2) is used in (8) as :

$$\hat{W} = \arg\max_{W} \{(1-\alpha)(\log[P(Y|F,W)] + \log[P(\omega)]) + \alpha(\log[P(f)])\} \qquad \textbf{(8)}$$

and it will adjust the final score to the Statistical Phonetic Modelling's Contribution.

4. RESULTS

The next table summarizes the results achieved in this research:

Recognizers	***%Correct***	***%Accuracy***
Standard Recognition	53,29%	13,31%
Statistical Phonetic Modeling Recognition	54,73%	17,31%
With 1 phoneme in the common sequence		
Herman Ney Standard Recognition	56,27%	15,30%
Herman Ney with Statistical Phonetic Modeling Recognition	59,50%	19,77%
With 2 phonemes in the Common sequence		
Herman Ney Standard Recognition	57,27%	18,50%
Herman Ney with Statistical Phonetic Modelling Recognition	61,14%	22,17%

Table (t-2): Results summary

[1] CLSP (Center for Language and Speech Processing), John Hopkins University, http://www.clsp.jhu.edu

[2] Because these terms are essentially the same just depending on the sort of approach considered: bottom-up or top-down.

[3] Again those terms are similar, depending on the adopted approach .

Four systems were prepared in order to evaluate the real improvement provided by the statistical phonetic modeling, using different algorithms.

First, a standard speech recognition system was created, using a classical Viterbi algorithm to score the maximum likelihood in the speech recognition process. This score will be our comparison reference with other results achieved in this work.

The second system is a standard speech recognition enhanced using the statistical phonetic modeling, re-scoring the likelihood computed in the testing step.

The third one is a standard Herman Ney system [13], using a Lexical Tree[4] with 1 (one) and 2 (two) phonemes in the common sequence, created from the same phonetic dictionary used in the standard recognition system.

The forth one is the same standard Herman Ney System created before (with 1 and 2 phonemes in the common sequence), enhanced now by statistical phonetic modeling, again re-scoring the results achieved in the recognition process.

5. CONCLUSION

The results point out that a statistical phonetic modelling reaches a significative increase in performance, considering that the SWITCHBOARD is a **spontaneous language database**, which acts as an additional difficulty term in the whole system's performance, as shown before in the table. Also, it should be pointed out that, the entire evaluation was under a speaker independent system scenario, which brings another variable to the system performance. This new proposed model shows a new reference for future work with spontaneous speech.

As a future work, this system could be re-evaluated , using another speech database (for instance the WSJ corpus) in training and testing steps, in order to reach a more solid comparison, that is, an increase in the % of acceptance and accuracy among other speech recognition systems. Until now preliminary results are very promising.

6. REFERENCES

[1] DELLER, J., PROAKIS, J.G., HANSEN, J. H.L. *Discrete-time processing of speech signals*. New York : Macmillan, 1993.

[2] ALLEN, J. Overview of text-to-speech systems. In: FURUI, S., SANDHI, M.M. *Advances in Speech Signal Processing.* New York : Marcel Dekker, 1992, p.741-790

[3] RABINER, L.R., JUANG, B.H. An introduction to Hidden Markov Models. *IEEE Acoustics, Speech, and Signal Processing*, p. 4-16, jan. 1986.

[4] RABINER, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, v.77, n.2, p.257-86, Feb. 1989.

[5] HUANG, X.D.; ARIKI, Y.; JACK, M.A. *Hidden Markov models for speech recognition*. Edinburgh, Edinburgh University Press, 1990.

[6] FAGUNDES, Rubem Dutra Ribeiro. *Reconhecimento de Voz, Linguagem Contínua, Usando Modelos de Markov*. São Paulo, 1993. Master's Dissertation - Escola Politécnica, Universidade de São Paulo. BRASIL

[7] NEY, H. et alii. Improvements in beam search for 10000-word continuous speech recognition. *IEEE International Conference on Acoustic Speech and Signal Processing*, p.I-9 - I-12, 1992.

[8] KENNY, P., HOLLAN, R., GUPTA, V.N., at al. A* - Admissible Heuristics for Rapid Lexical Access. *IEEE Transactions on Speech and Audio Proceedings*, vol.1, n.1, p.49-58, Jan. 1993.

[9] KENNY, P.; LI, Z., O'SHAUGHNESSY, D. Searching with a transcription graph. *IEEE International Conference on Acoustic Speech and Signal Processing*, p. 564-567, 1995.

[10] JELINEK, F.; BAHL, L.R.; MERCER, R.L. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, v.IT-21, n.3, p. 250-256, may 1975.

[11] ROUKOS S. Language representation. In: COLE, R. (ed.) *Survey of the state of the art in human language technology*, p.35-41, 1995. http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html,.

[12] LACOUTURE, R. *Au sujet des algorithmes de recherche des systèmes de reconnaissance de la parole à grands vocabulaires*. Thèse de Doctorat, McGill University, Montreal, 1995.

[13] NEY, H.; AUBERT, X. Dinamic programming search strategies: from digit strings to large vocabulary word graphs. In.: LEE, C.H. et alii (eds.) *Automatic speech and speaker recognition*. Boston : Kluwer Academic Publishers, 1996.

[4] Both lexical trees are generated by a new tool created specially for this work, named HLextree