Report on the paper SIGNALS TO SYMBOLS TO MEANING: MACHINE UNDERSTANDING OF SPOKEN LANGUAGE

Renato De Mori
McGill University, School of Computer Science
3480 University street - Montreal - Quebec - Canada

ABSTRACT
This paper is a report on V. Zue's invited paper : FROM SIGNALS TO SYMBOLS TO MEANING : ON MACHINE UNDERSTANDING OF SPOKEN LANGUAGE

1. CONSIDERATIONS ON THE PAPER'S CONTENT

The paper of Professor Zue is an accurate and concise review of the State of the Art in Automatic Speech Recognotion (ASR). Discussing results in terms of intervals of language model perplexities is interesting even if, as Zue points out, lower perplexities do not necessarily imply easier tasks.

Zue's reasons why there is a stochastic component that has to be added to speech and linguistic knowledge are also pertinent. In fact speakers may convey the same underlying message with many choices and linguistic constructs.

Another important aspect emphasized by Zue's paper is the need to compare experimental results by using public domain speech corpora.

Also of interest are considerations on the use of an ear model, feature extraction and the description of the Voyager system.

2. SOME ADDITIONAL CONSIDERATIONS ON HIDDEN MARKOV MODELS

Hidden Markov Models (HMM) have been very popular and highly successful tools for acoustic modelling. The following aspects seem to be of interest:

1) How many different models are required for recognizing large vocabularies?

Triphone models seem to be the solution adopted by many researchers, but their number is very high, making parameter estimation not very accurate with the available data. Various techniques for smoothing and clustering have been proposed (an interesting clustering algorithm has been recently presented by Bahl [1]). Interesting ideas have been proposed by Paul [2] and Bartakova and Jouvet [3] trying to take into account phonetic contexts inspired by phonetic knowledge in such a way that the number of units is kept in the order of magnitude of a few thousand. It is also important to note that in many recent systems triphone models are highly influenced by the word used to extract their parameters and tend to be word dependent.

2) Is corrective training a valid approach?

Recently, various types of algorithms have been developed for such a purpose, based, for example, on Maximum Mutual Information Estimation [4].

3) Are there new acoustic parameters worth using?

There seem to be a tendency of considering new dynamic parameters, like the second derivative of energy and mel-scaled cepstral coefficients [5]. The parameters and the approach for feature extraction mentioned by Zue in his review are also worth mentioning.

4) What is the role of Artificial Neural Networks (ANN)?

ANNs are essentially function approximators. They can approximate classifiers or data compressors that compute a reduced set of acoustic parameters based on a large set of parameters considered by a designer relevant for performing some acoustic classification. It is important, for this purpose, to have algorithms for global parameter estimation, i.e. for estimating at the same time the parameters of an ANN and the parameters of an HMM which takes the output of the ANN as an observation. Interesting algorithms can be found in [6-7].

ANNs have not succesfully modeled so far the dynamic characteristics of speech processes. Nevertheless, they have produced promising results as classifiers/detectors of some phonetic features. This possibility may be useful in future. A system could be conceived with a set of networks, each network being specialized in the detection of a complete set of features. The network specialized for a set can be fed by acoustic parameters suitable for them in terms of groups of features and their contexts, making different units only for those contexts for which feature variations imply important coarticulation effects.

3. LANGUAGE AND DIALOGUE MODELLING

In Automatic Speech Recognition, the words of a sentence are not available when the sentence has to be interpreted, or if they are available, they are affected by errors. This makes the understanding problem a search problem in which partial interpretation theories are scored. A coherent scoring methodology has been developed based on probabilities. So far linguistic knowledge has been mostly represented by trigram probabilities of having a word or a Part Of Speech (POS) given the two preceding ones. Estimating the probabilities required by these models presents some problems even if large corpora are available. These problems have found interesting solutions in recent years. Nevertheless, these models take into account only a limited context for a word, while it is well known that the expectation of a word in a sentence may depend on the entire sentence structure and, more generally, on the state of the conversation.

An interesting approach emerging now is based on stochastic grammars. Relevant problems along this line include for example the computation of the probability that a grammar generates a sentence only a part of which is known (see [8] and [9] for examples).

Another interesting problem is related to the opportunity of accepting only partial parses of a spoken sentence instead of forcing a complete parse of it.

The role of semantics is also of fundamental importance in speech decoding as a filter, as Zue pointed out, for example, in situations which a recognizer produces the N best word sequences (an efficient algorithm has been proposed recently for this purpose [10]) using only a language model based on bigram probabilities. Another important role of semantic knowledge could be that of predicting new words to be detected in the signal. For this purpose, new mathematical frameworks for Language Modeling have to be developed. A relevant problem is also that of linguistic knowledge acquisition from written corpora.

Finally there is an emerging interest in Dynamic Language Models in which the expectation of a word is considered as a function of the state of the verbal message or the dialogue. Cache memories can be used for this purpose [11].

4. REFERENCES

[1] L. R.Bahl, P. V. De Souza, P. S. Gopalakrishnan, D. Nahamoo and M. A. Picheny, Context-Dependent Modeling of Phones in Continuous Speech. Proc. 4th DARPA Speech and Natural Language Workshop, Asilomar, CA, Feb. 1991

[2] D. B. Paul, New Results with the Lincoln Tied-Mixture HMM-CSR System. ibid.

[3] K. Bartakova and D. Jouvet , Allophone Modelling 12th International Phonetic Congress, Aix en Provence, Aug. 1991

[4] Y. Normandin and S. Morgera, An Improved MMI Training Algorithm for Speaker-Independent, Small Vocabulary Continuous Speech Recognition. IEEE Intl. Conf. On Acoustics, Speech and Signal Processing, Toronto, Ontario, May 1991

[5]J. C. Wilpon, C. H. Lee and L. R. Rabiner, Improvemennts in Continuous Digit Recognition Using High Order Spectral and Energy Features. ibid.

[6] J. Bridle, A Recurrent Neural Network Architecture with a Hidden Markov Model Interpretation. Speech Communication, vol. 9, no. 1, Feb. 1990, pp.83-92.

[7] Y. Bengio, R. De Mori, G. Flammia and R. Kompe, Global Optimization of a Neural Netwotk-Hidden Markov Model Hybrid. Mc Gill University, School of Computer Science, Internal Report, Dec. 1990.

[8] F. Jelinek, Computation of the Probability of of Initial Substring Generation by Stochastic Context-Free Grammars. IBM Internat Report, April 1990.

[9] A. Corazza, R. De Mori, R. Gretter and G. Satta, Computation of Probabilities for an Island Driven Parsers. IEEE Transactions on Pattern Analysis and Machine Intelligence (to appear).

[10] F. K. Soong and E. F. Huang, A Tree-Trellis Based Fast Search For Finding N best Hypotheses in Continuous Speech recognition. IEEE Intl Conference on Acoustics, Speech and Signal Processing, Toronto, Ontario, 1991

[11] R. Kuhn and R. De Mori, A Cache-Based Natural Language Model for Speech Recognition IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 6, June 1990, pp. 570-583.