392a

Paper's Title:

An Auditory Neural Network for the Phonetic Encoding of Spoken Spanish

Authors:
M. Muñoz[1], C. García[1], A. Gómez[1], V. Rodellar[1], P. Gómez[1], R. W. Newcomb[2]

Affiliations:

[1] Depto. de Arquitectura y T. de S. I.
Facultad de Informática
Campus de Montegancedo, s/n
28660 Boadilla del Monte
Madrid - SPAIN

[2] Microsystems Lab.
Electrical Engineering Dept.
University of Maryland
College Park, MD 20742
U.S.A.

Paper's Abstract:

Introduction. One of the most difficult problems to be afforded in Natural Speech Recognition is that produced by the alteration of regular sounds by certain speakers. This produces a great variety of associated forms for a given speech utterance which renders the task of Speech Recognition a very difficult one. A special case present in both European and American Spanish, is the dilution of the intervowel voiced fricatives, as happens especially with /β/δ/γ/, when working in structures of the kind VCV. Many examples could be cited, the phenomenon being currently very active, associated with local dialectalism, the cultural background of the speaker, its animical condition, the speech rate, and other factors. The first objective of the research described in the present paper, is to study the relations and evolution of the phonemes /β/δ/γ/ and their plosive relatives /p/t/k/ and /b/d/g/ in groups VCV, to produce a Phonetic Coding Scheme, and to desing and build a Neural Network which could code and identify such sounds in a connected speech context. As a second objective, the Phonetic Coding Scheme will be extended to other consonants, to form a regular and orthogonal set of consonant sounds. Finally, provisions are made to include in the Phonetic Coding Scheme and in the Neural Network being trained the rest of the sounds of interest in Spanish in a step by step basis.

Correlation Measure among consonants in groups VCV. A preliminary study was carried out on the temporal evolution of the sequences of vectors derived from parameter extraction by LPC [Par.87, Sha.87], involving the sounds /p/t/k/b/d/g/β/δ/γ/ which will be resumed in the paper. This produces a first Phonetic Encoding Scheme, according with the Articulation Place, the quality of Voiced/Unvoiced, and the Degree of Closure. An example of the nine demisillables of the type CV containing the above group and the vowel /a/ were registered. Such frames were processed to detect the pitch, and divided into subframes containing one pitch period each. An LPC extraction based on Itakura's method [Par.87, Sha.87] was carried out on each frame, yielding an LPC vector of 16 coeficients per frame. Each of these LPC vectors was compared with the other 11 using a cosine measure [Tou.85]. From this study it was derived that there was a phonetic boundary in all the cases studied, marked by the existence of two highly correlated areas separated by a rapid change in the measure of correlation around a certain frame. A procedure for detecting the phonetic boundary between consonant and vowel may be implemented using a window of three frames of LPC vectors. The consonant can be identified from its spectral evolution [Par.87]. These facts were later used in the definition of the Neural Networks being designed, and will be given and discussed in the paper.

Extended Phonetic Coding Scheme. To produce an efficient and comprehensive Phonetic Coding Scheme is a crucial objective in the context of Speech Recognition by Neural Networks. We produced a bit-based encoding scheme, in which each independent Phonetic Feature would be encoded as follows:

Bit Position	Feature	Comment
0	Energy	Distinguishes among sound/silence
1	Voice	Distinguishes among voiced/unvoiced
2	Opening 1	Quantify the opening of the vocal
3	Opening 2	tract: 11-VW 10-SC 01-FC 00-PL
4	Articul 1	Quantify the place of articulation:
5	Articul 2	11-LP 10-TH 01-PT 00-VL
6	Special 1	VW: RN/OV; SC: LT/FT
7	Special 2	LT: Multiplicity; Others: Tension

VW: Vowel	LP: Lips	RN: Rounded
SC: Semiconsonant	TH: Teeth	OV: Oval
FC: Fricative	PT: Palate	LT: Lateral
PL: Plosive	VL: Veil	FT: Frontal

From this Phonetic Encoding Scheme we extracted a subset of phonemes considering only bits b_1, b_3, b_4 and b_5, which would code the sounds /k/, /c/, /t/, /p/, /ç/, /x/, /ʃ/, /θ/, /s/, /φ/, /f/, /g/, /j/, /d/, /b/, /γ/, /ζ/, /δ/, /z/, /β/. A graphical representation could be assigned to this Phonetic Coding Scheme in the shape of a hypercube. This structure reveals a way to design and train a Neural Network to implement the separation among the different sounds, which will be presented and discussed.

System under design. The global system to be put up to work may be seen in Fig. 1. The trace of speech has been sampled yielding a sequence of numbers x(n). This signal is divided into overlaping frames of 128 samples, which are fed to an LPC extractor. The output of the extractor is a vector of dimension 16, $\{e_k\}$, which will be phonetically encoded in the block TDNN yielding $\{c_m\}$, which could be used different purposes in later processing.

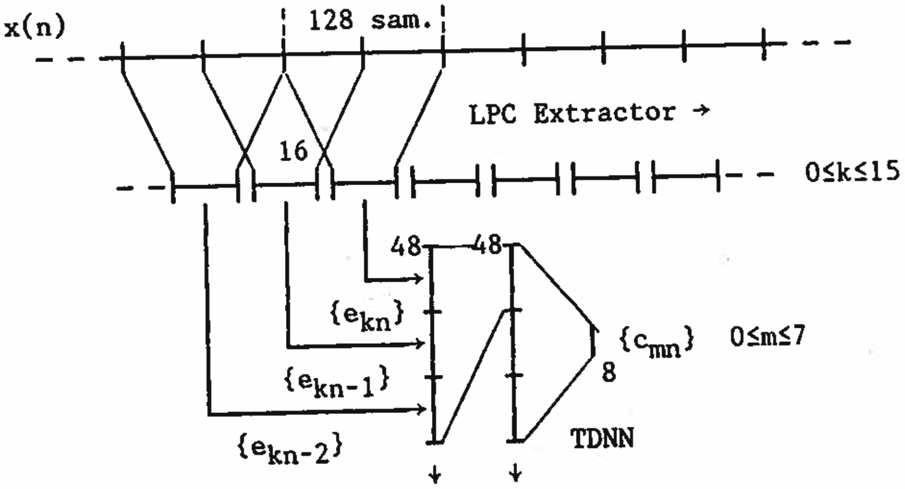

Fig. 1. LPC Extraction and TDNN Phonetic Coding.

A Time Delay Neural Network (TDNN) [Wai.89] is a Back Propagation Neural Network [Kha.90] in which the vectors processed in each layer are time delayed copies of themselves. Another approach, which is being studied, makes use of an Auditory Model [Gom.87, Gom.90] to produce a vector with temporal and spectral information to be supplied to the TDNN. In this case, the size of the vector of parameters could be reduced down to 6.

Simulation and Implementation. Several lines are being followed to carry the above models into practice, which will be discussed in brief in the paper. At a first level, a simulation process of the whole system is being carried out on a SUN Sparc Workstation. A simulated version of the TDNN being designed have been developed, and a Training System and a Training Set have been put to work. The whole system is being conceived for its operation in real time on more specialized processors, for which a simulation is being conducted on a DSP32C Digital Signal Processor, to build a more compact subsystem amenable of being devoted to real applications of phonetic encoding, such as Assisted Foreign Language Learning and others. Finally, a VLSI Chip is being designed to serve as a specialized Phonetic Encoder, comprissing LPC or Auditory Parameter Extraction and TDNN coding. The chip will provide interfacing for its systolic interconnection. For this reason a research to study the systolization of the algorithms being designed is also being conducted on a Supernode System provided by a Parallel Computing Action promoted by the ESPRIT Initiative within the Research Programs of the European Community.

Acknowledgements. The present work and its related activities are being conducted under the support of Acción Especial MIC88-0398-E from the Spanish Government Plan Nacional de Microelectrónica, the Parallel Computer Action Initiative of the ESPRIT Program, and the donation of DSP32 equipment by AT&T Microelectronics, Spain.

References.

[Gom.87] "VLSI Implementation of a Digital Filter for Hearing Aids", P. Gómez, V. Rodellar, M. Hermida, A. Díaz and R. W. Newcomb, Proc. of the 1987 Digital Signal Processing Conference, Ed. V. Capellini and A. G. Constantinides, Florence, Italy, 1987, pp. 341-345.

[Gom.90] "Métodos Numéricos en el Desarrollo de Filtros Espacio-Temporales en Percepción Auditiva", P. Gómez, V. Rodellar, A. Díaz and M. Hermida, Memorias del I Congreso en Métodos Numéricos en Ingeniería, Ed. G. Winter and M. Galante, Las Palmas de Gran Canaria, Spain, 1990, pp. 249-256.

[Kha.90] "Foundations of Neural Networks", T. Khanna, Addison-Wesley, Reading, Massachusetts, 1990.

[Par.87] "Voice and Speech Processing", T. Parsons, McGraw-Hill, New York, 1987.

[Sha.87] "Speech Communication: Human and Machine", D. O'Shaughnessy, Addison-Wesley, Reading, Massachusetts, 1987.

[Tou.85] "Pattern Recognition Principles", J. T. Tou and R. C. Gonzalez, Addison-Wesley, Reading, Massachusetts, 1985.

[Wai.89] "Phoneme Recognition Using Time-Delay Neural Networks", A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. J. Lang, IEEE Trans. on ASSP, Vol. 37, No. 3, March 1989, pp. 328-339.