Using Neural Networks in the Estimation of Consonant Imprecision Ratings.

Carlos A. Ferrer, Anesto del Toro, Eduardo González & Maria E. Hernández-Díaz.

Center for Studies on Electronics and Information Technologies, Central University of Las Villas, Zip code: 54830, Santa Clara, Cuba.
{cferrer, anestodt, moreira, mariae}@uclv.edu.cu

Abstract. This paper deals with the objective measurement of the articulatory imprecision severity in patients with motor speech disorders. Acoustic recordings of repetitive sequences of plosive consonants from 58 patients are used, along with the corresponding subjective ratings of consonant imprecision made by two judges. An estimate of the subjective perception of imprecision per patient is made using energy and sonority measurements of the acoustic signal. Several Neural Networks (NN's) architectures are tested and compared to the linear regression approach to evaluate their prediction abilities. A reduction of more than 25% of the linear regressions error variance is obtained with the use of NN's without a significant increment of computational requirements.

1 Introduction

There is a general agreement that objective measures should be used in the assessment of voice disorders, as a complement to the perceptual judgments of the specialist [1][2][3][4]. Features related to voice quality, pitch perturbations and laryngeal function have been widely addressed, and acoustic correlates of them have been devised. In other symptoms, like those related to prosody, articulation, and nasality fewer results have been accomplished. This paper proposes a method to obtain an estimate of the perceived articulatory imprecision in voiceless plosive consonants. To this end, measures of energy and sonority are used, and several function approximators are tested, including multiple linear regression and feed-forward neural networks.

The structure of this paper is as follows: Section 1 introduces the antecedents of this work, regarding the methods to obtain the measurements of energy and sonority used as inputs to the approximators. In Section 2 the selection of the neural network topologies employed is discussed, together with the experiments conceived to evaluate their effectiveness. In Section 3 the results of the experiments are shown and analyzed, focusing on network performance and generalization capacity. In Section 4 the conclusions of this work are stated, along with recommendations for further research.

The following subsections introduce the reader with the motives and explanation of some approaches that might seem otherwise arbitrary.

1.1 Perceptual Analysis

The term motor speech disorders (MSD) stands for pathologies that cause a disturbance in the control of speech muscular movements as a consequence of a lesion in the central or peripheral nervous system. There are two types of MSD: dysarthria and apraxia. Due to their characteristics [5], it is in dysarthria and not in apraxia where there can be consistent acoustic correlates of the perceived characteristics. Some examples of dysarthrias are Parkinson Disease, Chorea, Amiotropic Lateral Sclerosis.

Several studies [5][6][7] carried out in the late 60's and early 70's were focused on the perceptual characteristics of dysarthric speech. The results of these and other related studies are still considered [4] the basis of clinical differential diagnosis of dysarthria. The methodology created consisted in the realization of three exercises by the patient, and the judgment of 38 perceptual characteristics by a panel of three judges. The exercises proved to convey the maximum clinical information in the minimum time possible [5], and consisted in:

- The phonation of a sustained vowel ("a"), that allows the panel to judge the quality, amplitude, duration and persistency of the fonatory control.
- The repetition of series of syllables using plosive consonants ("Pa" "Ta" and "Ka"), as fast and steady as possible, giving information of rhythm, regularity and duration of every kind of articulatory movements.
- The reading of a standard paragraph, to appraise the way the patient integrates the phonatory, resonatory and prosodic characteristics of contextual speech.

The 38 perceptual characteristics were judged in a 7 point scale, from 0 to 6, with 0 the least perceivable and 6 the most severe level of the feature.

It was shown in these studies that each dysarthria is described by a unique set of groups (clusters) of perceptual characteristics, and that differential diagnosis of dysarthria is possible on the basis of the way the speech sounds.

1.2 Objective Measurement of Consonant Imprecision

This paper is focused on the objective measurement of consonant imprecision, one of the 38 dimensions used in [6][7]. Consonant Imprecision was found significant in all the types of dysarthria reported in the mentioned studies. This fact makes a measurement of consonant imprecision non useful in differential diagnosis, but a good indicator instead of therapy adequacy and for rehabilitation documentation in most dysartrias. To the authors' knowledge, there are no previous objective indexes reported to quantify the degree of consonant imprecision, but only indexes related to the percentage of correctly perceived consonants in a predefined word set [8]. In previous studies by the authors [9][10], it was decided to work with recordings of the Pa/Ta/Ka exercise to obtain an objective measurement of consonant imprecision for several reasons:

- This exercise is part of the standard Mayo Clinic Methodology.
- It is used to appraise the articulatory functioning, where consonant imprecision is included.

- The determination of the position of the consonants can be located more easily than in the paragraph reading.
- All the consonants have similar characteristics (voiceless stops), making easier to devise a procedure for measuring deviations from normality.
- For the latter reason, there is no need to detect the particular consonant of each syllable.

The common characteristics of the consonants present in the Pa/Ta/Ka are the release of a total occlusion in a certain place of the vocal tract, and the lack of sonority previous to that release. A listening to the Pa/Ta/Ka recordings revealed that the possible distortions present were sonorization, nasalization and affrication (see Table 1). The distortions found were the presence of energy prior to the release of the occlusion (turbulent noise in fricatives and periodic sounds in nasal and voiced consonants) and the presence of periodicity before the release in nasal and voiced consonants. Two indexes of abnormality were then established, related to the energy and the level of sonority prior to the release of the constriction.

Table 1. Substitutions found in Pa/Ta/Ka recordings, grouped by the place of the constriction in the vocal tract

	Labial	Palatal	Velar
Unvoiced Stops (Original)	Pa	Ta	Ka
Voiced Stops	Ba	Da	Ga
Nasal (Voiced)	Ma	Na	-
Fricatives (Unvoiced)	Fa	Sa	Ja

The first index, denoted *CIE* (*C*onsonant *I*mprecision by *E*nergy), is a ratio of the areas of the energy envelope before (*Ab*) and after (*Aa*) the release of the constriction (see Fig. 1) as defined in equation (1).

$$CIE = \frac{A_b}{A_a} = \frac{\sum_{T_{on}-T_r}^{T_{on}-T_r} E(t)}{\sum_{T_{on}+T_r}^{T_{on}+T_r+T_a} E(t)} \qquad (1)$$

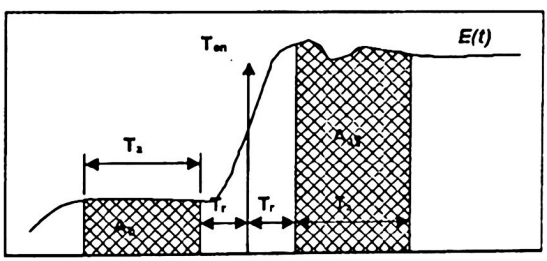

Fig. 1. Graphical representation of the variables used in the calculation of *CIE*.

The instant of release is denoted T_{on} in (1), and the values of the time intervals T_r and T_a were 10 and 20 milliseconds (*ms*), respectively. The average of the *CIE* values

for all the syllables found in the patient's recording (CIE_m) was used as the objective index of consonant imprecision regarding inadequate release of the occlusion.

The second index was calculated using the autocorrelation function of a segment of the speech signal previous to the T_{on}. The value of the maximum peak of the autocorrelation in the range of possible values of fundamental period of voice (2-20 *ms*, 50-500 Hz) is found on each syllable. The average of these maximums is denoted CIS_m (*C*onsonant *I*mprecision by *S*onority *m*ean) and is calculated for each patient's recording. The speech segment's end is set 10 *ms* before T_{on}, and its length is 40 *ms*.

For the calculation of both indexes, the determination of T_{on} is required and crucial. To this end, a syllable detector was devised [9], based on the analysis of the energy envelope maximums. The maximums that do not satisfy some heuristics that must be met to be perceived as a separate syllable are eliminated from the list of syllable candidates, and the remaining maximums are considered syllable centers. The heuristics used were:

- Amplitude greater than 20 times the minimum value of the energy envelope.
- Separation between maximums greater than 100 *ms* (when two maximums are closer than this value, the one with the lower amplitude is suppressed).
- Presence of a minimum of less than 75% of the maximum's amplitude between itself and the previous and posterior maximums.
- Separation of the mentioned minimums of more than 50 *ms*.

Once the syllables are detected, the instant of release is determined as the point in the energy envelope with the greatest positive slope in the segment between the syllable's center and the previous one.

This method of syllable position determination showed an 89% of correct detection in a set of 3750 syllables of dysarthric patients [9].

With the obtained values of CIE_m and CIS_m and the subjective evaluations of two judges, the linear regression of the averaged subjective judgments (SJ_m) was obtained for both indexes and their combination. The results of the correlations and error variances of the three regressions to the original subjective judgments are shown in Table 2. The correlation between judges was 0.75.

Table 2. Correlations and error varainces of the linear regressions of CIE_m, CIS_m and their combination $CIES_{lr}$

	CIE_m	CIS_m	$CIES_{lr}$
Correlation with SJ_m	0.5866	0.5642	0.6711
Error Variance	2.3527	2.4451	1.9712

1.3 Hypotheses and Objectives

Regarding consonant imprecision, the analysis of the location in the CIS_m/CIE_m plane of the four types of consonants involved suggests that a linear relationship can be a gross estimate (See Fig. 2). It is considered by the authors that a linear regression of the indexes devised is not a precise approximation to the way human ears perceive consonant imprecision. If the subjective judgment is zero for "Normal" and six for the other three consonants, it is evident that a nonlinear approximation should outperform the linear regression approach.

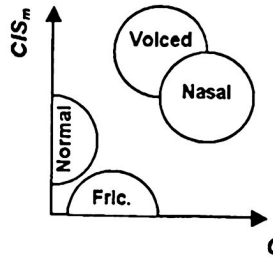

Fig. 2. Representation of the consonants in Table I on the plane CIS_m/CIE_m.

The objective of this paper is to explore the improvement in performance (i.e. reduction in error variance or increment in correlation) a nonlinear "approximator" (the term "predictor" will also be used in this paper) can achieve with no significant increment in model complexity compared to the multiple linear regression.

2 Non-Linear Models and Experiments.

Different non-linear prediction models, based on feed-forward neural networks (FNNs), were selected and used for testing purposes. The models included various topologies of multilayer perceptrons (MLP) and radial basis networks (RBN) [11]. These architectures had been widely used and their abilities as function approximators had been demonstrated in different applications [12].

2.1 Topology Selection.

The maximum number of neurons for the FNNs was set to 5. This was decided to avoid a significant increment in the predictor's model complexity compared to the linear regression approach.

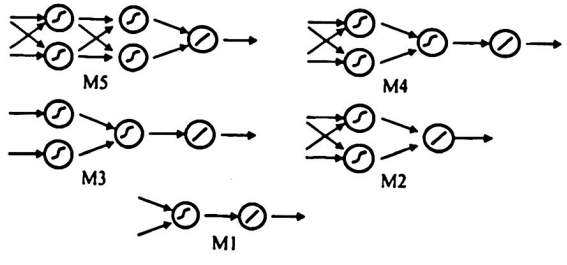

Fig. 3. MLPs' topologies tested.

For the MLP architecture the topologies used are shown in Figure 3, in descending order of complexity. From now on they will be referred as M5, M4, M3, M2 and M1, respectively. A linear positive transfer function was employed in the output layer while the hyperbolic tangent sigmoid was used for the input and hidden layer neurons [11]. The RBF networks evaluated ranged from 2 to 5 neurons, and will be referred correspondingly as R2, R3, R4 and R5.

2.2 Training Experiments

The data used for the evaluation of the predictors were the obtained CIE_m and CIS_m indexes for 58 recordings of the Pa/Ta/Ka exercise, along with the corresponding subjective ratings of consonant imprecision taken from [10]. In all cases the objective indexes were employed as inputs and the subjective ratings were used as output targets.

The generalization ability of each topology was tested (experiment "A"), to have a better appraisal of the tradeoff with the predictor's performance. To this end, the whole data set was randomly divided in two halves, a training and a control set. Each MLP was randomly initialized several (15) times to reduce the chance of convergence to local minimums, a common failure of the *backpropagation* learning function used, and the whole training set was fed to the network 300 times, when the training process was stopped. The best resulting MLP among the 15 initializations was chosen as the optimal predictor for the training set. The RBF networks were obtained for the same training set, with different spreads of the radial basis function. The best resulting RBF among the different spreads was considered the optimal approximator within each topology. This halving procedure was repeated 60 times, and the averaged results for training and control sets for both MLP and RBF networks were obtained. Each network was also trained for the whole data set (experiment "B") to obtain their global performances, using the same random initialization procedure for the MLPs.

3 Results and Discussion

The performances of the networks for the total data set and the training/control experiment are shown in Tables 3 (RBFs) and 4 (MLPs and linear regression). Table 4 includes the number of outliers obtained in experiment "A" for the 60 control sets. An error variance greater than ten times the median of the results with the control sets was considered as outlier. The mean error variance of the control sets results with the outliers removed is also shown in Table 4 (row denoted "C.S. w/o Outl."), since the one including the outliers is practically meaningless. Another extra row in Table 4 is the value of the Pearson's correlation coefficient obtained between the subjective ratings and the outputs of the MLPs predictors in experiment "B".

Table 3. Mean error variances of the 60 training/control experiments and total error variance for the RBF networks

		R2	*R3*	*R4*	*R5*
A)	Training Set	2.2354	1.7110	1.6478	1.5734
	Control Set	2.5247	2.2109	2.2855	2.7619
	Average	2.3801	1.9609	1.9667	2.1677
B)	Whole Set	2.4794	2.211	1.7531	1.7416

Table 4. Mean error variances of the 60 training/control experiments and total error variance for the MLP networks and the Linear Regression approach.

		M5	*M4*	*M3*	*M2*	*M1*	*LR*
A)	Training Set	0.9893	1.1126	1.3870	1.2427	1.7779	1.7615
	Control Set	*66.643*	*20.418*	*2.4949*	*2469.4*	*2.4995*	*2.2449*
	# of Outliers	*8*	*2*	*0*	*1*	*0*	*0*
	C.S. w/o Outl.	4.5264	3.2534	2.4949	2.8237	2.4995	2.2449
B)	Average	2.4977	2.1649	1.9404	2.0266	2.1387	2.0052
	Whole Set	1.3940	1.5136	1.6938	1.5239	1.9163	1.9712
	R	0.7819	0.7603	0.7265	0.7584	0.6824	0.6711

From these results it is apparent than RBFs of up to 5 neurons are not well suited to be good approximators for this data set. The best overall performance (obtained for R5) only surpasses the linear regression and M1 results. Besides, there is no significant increment in performance from R4 to R5, while the generalization ability is actually deteriorated. The lower the average performance of a topology in experiment "A" the higher its generalization ability.

The MLP architecture showed a better performance, although some topologies presented outliers in the results. There is an exception in the logical increment sequence of performance in experiment B) going from topology M1 to M5. It is in the case of M2 and M3, where the latter has one neuron more than the former, and in spite of this, M3 has a lower performance and a better generalization capacity. This is due to the greater number of connections used by M2, allowing the formation of more complex surfaces than M3. It is precisely this pair of topologies the ones with better results taking into account global performance (in terms of error variance), generalization capability and absence of outliers. The M2 topology has an error variance comparable to the ones of M4 and M5, together with the second best generalization ability, and only one outlier result, while M3 has no outlier, presents the best generalization ability, but shows a lower global performance. Even though, the correlation coefficient for the M3 predictor is higher than 0.707, so this model explains more than 50% of the error variance, result that is considered acceptable in the literature [13][14] when dealing with subjective judgments. The lowest error variance obtained (1.39) represents a reduction of almost 30% of the obtained with the linear regression (1.97). The values of correlation obtained (0.72~0.78) are comparable to the 0.75 between judges.

4 Conclusions.

From the two architectures of nonlinear approximators tested, the RBF networks did not show a good performance or generalization ability. The MLP predictors presented the best results, with a reduction of more than 25% of the error variance of the original linear regression approach in the 5 neurons topology. The best results considering all the factors evaluated, were obtained for the M2 and M3 variants, with 3 and 4 neurons, respectively. This represents no significant increment in the approximator model complexity compared to the linear regression. The values of correlations found are similar to the ones obtained between the subjective judgments of the specialists.

Further research is needed to obtain a third index in order to increase predictor's performance. Specifically, an index that could make a better separation of the fricatives from the normal plosive consonants (see Fig. 2.) would be desired.

References

[1] Baken, R.J. Clinical Measurement of Speech and Voice. Singular Publishing Group Inc. San Diego. (1996)

[2] Cannito, M.P., Yorkston, K.M. & Beukelman, D.R. Neuromotor speech disorders. Nature, Assessment, and Management. Paul H. Brookes Publishing Co. (1998).

[3] Kent, R.D. & Ball, M.J. Voice Quality Measurement. Singular Publishing Group, Inc. (2000).

[4] Yorkston, K.M., Beukelman, D.R. & Bell K.R. Clinical Management of Dysarthric Speakers. Austin, TX: Pro-Ed. (1988).

[5] Darley, F.L.; Aronson, A.E. & Brown, J.R. Motor Speech Disorders. Philadelphia. Saunders. (1975).

[6] Darley, F.L.; Aronson, A.E. & Brown, J.R. Clusters of deviant speech dimensions in the dysarthria. Journal of Speech & Hearing Research. 12, pp 462-496. (1969).

[7] Darley, F.L.; Aronson, A.E. & Brown, J.R. Differential diagnostic patterns of dysarthria. Journal of Speech & Hearing Research. 12, pp 246-269. (1969).

[8] House, A. S. et al, Articulation-Testing Method: Consonantal Differentiation with a Closed-Response Set J. Acoust. Soc. Am. Vol 37 (1), pp 158-166. (1965).

[9] Ferrer, C.A.; Hernández, M.E. & González, E. Isolated Syllable Position Detector in Recordings of Patients With Motor Speech Disorders Using Speech Processing Techniques. Proceedings of the TELEC'02 International Conference, Santiago de Cuba, ISBN 84-8138-506-9, July. (2002).

[10] Ferrer, C.A. & Hernández, M.E. A Measure of Articulatory Imprecision in Dysarthric Patients Recordings. Proceedings of the VIII International Congress on Social Communication. Santiago de Cuba, ISBN 84-8138-506-9, January. (2003).

[11] H. Demuh, Neural Network Design, PWS Publishing Company. (1996).

[12] B. Widrow, E. Rumelhart, A. Lehr, Neural Networks: Applications in industry, business and science, Communication of ACM, 37 (3) 93-105. (1994).

[13] Rabinov C. R., Kreiman J. Comparing Reliability of Perceptual Ratings of Roughness and acoustic Measures of Jitter. Journal of Speech & Hearing Research. 38, pp 26-32. (1995).

[14] Fukazawa, T.; El-Assuooty, A. & Honjo, I. A new index for evaluation of the turbulent noise in pathological voice. Journal of the Acoustical Society of America. Vol. 83, No 3. pp 1189-1193. March. (1988).