Voice Signal Typing Using a Pattern Recognition Approach☆

[*]J.M. Miramont, [*]Juan F. Restrepo, [†]J. Codino, [†,‡]C. Jackson-Menaldi, and [*]G. Schlotthauer, *Oro Verde, Argentina, and †Clair Shores, and ‡Detroit, Michigan

Abstract: Voice signal classification in three types according to their degree of periodicity, a task known as signal typing, is a relevant preprocessing step before computing any perturbation measures. However, it is a time consuming and subjective activity. This has given rise to interest in automatic systems that use objective measures to distinguish among the different signal types. The purpose of this paper is twofold. First, to propose a pattern recognition approach for automatic voice signal typing based on a multi-class linear Support Vector Machine, and using rather well-known parameters like Jitter, Shimmer, Harmonic-to-Noise Ratio, and Cepstral Prominence Peak in combination with nonlinear dynamics measures. Two novel features are also proposed as objective parameters. Second, to validate this approach using a large amount of signals coming from two well-known corpora using cross-dataset experiments to assess the generalizability of the system. A total amount of 1262 signals labeled by professional voice pathologists were used with this purpose. Statistically significant differences between all types were found for all features. Accuracies over 82.71% were estimated in all intra-datasets and inter-datasets using cross-validation. Finally, the use of posterior probabilities is proposed as a measure of the reliability of the assigned type. This could help clinicians to make a more informed decision about the type assigned to a voice. These outcomes suggest that the proposed approach can successfully discriminate among the three voice types, paving the way to a fully automatic tool for voice signal typing in the future.

Key Words: Voice signal typing−Voice signal classification−Support vector machine−Pattern recognition.

INTRODUCTION

Voice signals can be classified in three types, namely type 1, type 2, and type 3, according to their degree of periodicity. This task, known as signal typing, has become a relevant preprocessing step to determine the suitability of a signal to perturbation analysis, which depends heavily on the assumption that the assessed signal is nearly-periodic.[1]

Type 1 signals are indeed defined as nearly-periodic. In contrast, type 2 signals are defined as those with strong modulating and subharmonic frequencies, and type 3 signals are defined as those which lack an apparent periodic structure. This classification scheme is useful to prevent voice pathologists from using analysis tools that might be unreliable for unsuited signals, which could lead to misleading values that do not carry any physiological meaning.[2]

Based on the stated definitions, only type 1 voices can be reliably analyzed with perturbation measures, like Jitter and Shimmer estimates, although there is evidence that some type 2 voices might be meaningfully analyzed as well.[3,4] Considering this, many authors began to apply this classification scheme as a standard procedure.[5−12]

A fourth type of voice has been introduced by Sprecher et al.[13] In this new four-type classification scheme, types 1 and 2 keep their definitions unaltered, but now type 3 comprises only those voices that exhibit a chaotic behavior while type 4 comprises signals with a strong stochastic component. Both types lack an apparent periodic structure. This poses the problem of distinguishing between type 3 and type 4 voices, which is analogous to the problem of differentiating a chaotic (deterministic) dynamic from a stochastic (nondeterministic) one. Although a time series coming from a chaotic system and a purely stochastic time series might look the same to an observer, voice pathologists in Sprecher et al[13] seemed to be able to determine which signals had a deterministic structure (labeled as type 3) and which ones were stochastic (labeled as type 4). We believe that differentiating type 3 from type 4 signals is a fairly difficult task for a human observer.[4] In consequence, a high dimensional chaotic signal, which is type 3 by definition, could be wrongly labeled as a type 4. Moreover, the definition of type 4 voices is based on concepts coming from the nonlinear dynamic systems theory framework. For that reason, we will use the original three-type classification scheme throughout this paper.

Spectrograms are the most widely used tool for classifying voice signals in the different types, as proposed in Titze et al and Sprecher et al,[1,13] regardless of whether the three-type or the four-type scheme is employed. To complement the visual information given by the spectrograms, voice recordings can also be heard by the clinicians to detect noise or important pitch changes. In addition, plots of sound pressure vs time can be used to directly assess the signal periodicity. Since clinicians tend to evaluate diverse aspects of the voice by these perceptual and subjective means, signal typing has become a time-consuming and biased task that can be affected by external factors like evaluator's mood, experience, or professional background (ie, otolaryngologists, speech-language pathologists, etc).[14]

Accepted for publication March 26, 2020.
☆Conflicts of Interest: The authors declare that they have no conflicts of interest.
From the *Instituto de Investigación y Desarrollo en Bioingeniería y Bioinformática, UNER-CONICET, Oro Verde, Entre Ríos, Argentina; †Lakeshore Professional Voice Center, Lakeshore Ear, Nose and Throat Center, St. Clair Shores, Michigan; and the ‡Department of Otolaryngology, School of Medicine, Wayne State University, Detroit, Michigan.
Address correspondence and reprint requests to J. M. Miramont, Instituto de Investigación y Desarrollo en Bioingeniería y Bioinformática, CONICET Santa Fe, Oro Verde 3100 Entre Ríos, Argentina. E-mail: jmiramont@conicet.gov.ar
Journal of Voice, Vol. ■■, No. ■■, pp. ■■−■■
0892-1997

https://doi.org/10.1016/j.jvoice.2020.03.006

This has given rise to interest in systems that might assist clinicians in perceptual tasks, such as voice signal typing, and that use a pattern recognition approach, based on objective quantitative measures, to automatically classify signals. This could help voice pathologists to decrease the mentioned bias as well as the time devoted to this preprocessing step.[15,16]

Despite the fact that many quantitative measures have been described to determine a voice's type,[16–22] the classification step was not addressed directly in previous works (except Lee[16]). Furthermore, the suggested parameters were validated over a small amount of signals coming from the same corpus (the total number of signals ranged from 135 to 148). This number of signals might be insufficient to assess the potential of those methods in a real-world situation.

Based on the aforementioned, the aim of this paper is two-fold. First, to propose a pattern recognition approach for automatic signal typing based on objective measures and using rather simple linear classifiers. We shall introduce two novel parameters, called Normalized Principal Component Variance (NPCV) and Standard Deviation of the Normalized Principal Component Variance (SDNPCV) that can be used as objective measures for signal typing, along with other well-known features like Jitter, Shimmer, Harmonic-to-Noise ratio (HNR), Cepstral Prominence Peak (CPP), and nonlinear dynamics measures. Second, to validate this approach using a larger amount of signals coming from two well-known corpora using cross-dataset experiments.

The novelty of our approach is based on the following aspects: (1) Two new features are proposed as objective measures for signal typing that can complement other commonly used parameters. (2) A multiclass classifier is used and posterior probability outcome is proposed as a confidence measure on the predicted type. (3) Performance is estimated using two different corpora and more than 1200 signals, which aims to assess the generalizability of the system by cross-datasets experiments. (4) In contrast to other works[13,17–19,21] computation of nonlinear dynamics features was done by means of the recently proposed U-Correlation integral,[23,24] which is a reliable user-independent method.

The rest of this article is organized as follows. Section Materials and methods describes the used signals, the proposed features, and the classification procedure. Section Results shows the results obtained for the classification and the statistical analysis. Finally, results are summarized and discussed in Section Discussion, followed by conclusions in Section Conclusions.

MATERIALS AND METHODS

Datasets

The signals used in this work corresponded to a sustained /*a*/ phoneme, from subjects with pathological phonation and from the following datasets:

TABLE 1.
Voice Types Distribution For Both Corpora

	Type 1	Type 2	Type 3	Total
SAAR	175	330	108	613
MEEI	185	335	129	649

Massachussets Eye and Ear Infirmary (MEEI)

The Massachusetts Eye and Ear Infirmary Voice Disorders Database (MEEI),[25] distributed by Kay Elemetrics, is a well-known corpus within the voice processing community. It is comprised by approximately 750 signals corresponding to normal and pathological phonation, which were recorded with sampling frequencies of 25 kHz (pathological phonation) and 50 kHz (normophonic) and 16 bits of resolution. Signals from this dataset are already preprocessed in order to ensure the stable part of the phonation is present.

Saarbrücken voice database (SAAR)

Recorded by the Institut für Phonetik at Saarland University and the Phoniatry Section of the Caritas Clinic St. Theresia in Saarbrücken, Germany, this corpus[26] (SAAR) contains more than 2000 audio recordings from German speakers with normal and pathological phonation. They were recorded using a sampling frequency of 50 kHz and 16 bits of resolution. Signals from SAAR dataset were also preprocessed in order to decrease sampling frequency to 25 kHz and ensure a duration of 800 ms, deleting the onset and offset parts, in order to match their characteristics to those of the MEEI corpus.

Signals from both corpora with interruptions or other problems, such as the presence of other voices during the register or a duration under 800 ms, were discarded.

Finally, all the voices were labeled as type 1, 2, or 3 by experienced clinicians (third and fourth authors) using spectrograms and the other aforementioned methods for signal typing. Only those signals for which there was agreement between the observers over the assigned type were included in this study. Those labels were then used as the ground truth to train the classification models. Table 1 shows the voice types distribution.

Features

Temporal regularity of the voice signal might be broadly explained by two dimensions: periodicity and wave shape. Moreover, voice signal periodicity can be affected by some, or a combination of, the following factors: acoustic properties, like Jitter and Shimmer, and aspiration noise.[27] Because of the many factors involved, it has been impossible so far to come up with a single parameter that is able to objectively distinguish among all types of voices. The reason behind this might be that, like many other perceptual tasks, signal typing is a multidimensional problem.[15] Bearing this in mind, we shall propose a *combination* of features with the

aim of reflecting the different influences affecting voice signals' periodicity.

First, in order to quantify the amount of Jitter and Shimmer of a voice, we will use local Jitter and local Shimmer measures. Although several estimators of these acoustic properties have been proposed, those two are among the most commonly used perturbation measures. Local Jitter is the average absolute difference between consecutive periods duration divided by the average period duration. Local Shimmer can be described in a similar way but using the absolute difference between the maximum peak-to-peak amplitude of consecutive periods and the average peak-to-peak amplitude among all cycles.

Second, we shall focus on descriptors that measure the noise component that affects signal periodicity. A parameter directly related with this is Harmonic to Noise Ratio (HNR)[28] which quantifies the ratio between the energy of the harmonic component of the signal and the energy of the aperiodic component, like aspiration noise. Hence, HNR will take large values for very periodic signals and low values for signals with poor periodicity, making it a promising feature for signal classification.

Also related to periodicity is the Cepstral Peak Prominence (CPP),[29] which is the amplitude (in dB) of the first cepstral peak measured from a linear approximation of the noise level at the peak quefrency. This cepstral measure is highly correlated with the perceptual impression of breathiness, with general disphonia[27] and with the three types of voices used throughout this paper.[30] Since the perceptual cues are very important for voice pathologists during voice signal typing, it seems natural to consider this kind of features. An advantage of this measure is that, in contrast to perturbation measures, it needs neither cycle-by-cycle segmentation nor accurate fundamental frequency estimation.

So far, features that describe signal's periodicity have been presented, but little has been said about wave shape assessment. Although local Shimmer describes average changes in the waveform, it only represents changes in the greatest amplitude. In consequence, this measure is "blind" to more general changes of the overall morphology of the cycles. With the aim of quantifying those changes, we propose two new traits called Normalized Principal Component Variance (NPCV) and Standard Deviation of the Normalized Principal Component Variance (SDNPCV).

In order to compute NPCV, the signal must be framed into its individual cycles first. Then, all cycles are normalized in amplitude, for instance between 0 and 1, and in length, interpolating shorter cycles in order to match their length to that of the longest period. We shall call $\boldsymbol{p_i}$ to the ith normalized period of the signal. After this, all $\boldsymbol{p_i}$ are arranged as columns of a matrix $\boldsymbol{P}$ and the covariance matrix of $\boldsymbol{P}$ is computed along with its eigenvalues λ_i. NPCV is finally defined as:

$$NPCV = \frac{\lambda_{max}}{\sum_i \lambda_i} \tag{1}$$

where λ_{max} is the largest eigenvalue.

This is equivalent to the variance explained by the principal component of the set of periods of the signal. The lower the value of NPCV, the larger is the variability of the signal's waveform. Consequently, NPCV can be regarded as a measure of the signal's waveform variability.

Another feature was computed based on a short-time version of NPCV, which we have called Standard Deviation of Normalized Principal Component Variance (SDNPCV). In order to compute SDNPCV, the signal is segmented using rectangular windows of 100 ms and NPCV is estimated segment-wise using Equation 1. Then, dNPCV_m is computed for the mth segment as $\mathrm{dNPCV}_m = \mathrm{NPCV}_{m+1} - \mathrm{NPCV}_m$, ie, the difference between the value of NPCV of consecutive segments. Finally, SDNPCV is computed as:

$$\mathrm{SDNPCV} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M-1}\left(\mathrm{dNPCV}_m - \overline{\mathrm{dNPCV}}\right)^2}, \tag{2}$$

where M is the number of segments and dNPCV is the mean value of dNPCV_m for all signal frames. SDNPCV behaves in an opposite way to NPCV for each type of signal, being higher for type 3 and successively lower for types 2 and 1.

Lastly, we shall describe the nonlinear dynamic measures used in this article. These features have been previously used for signal typing.[13,17,19,21,23,31] They describe the dynamics of a time series and the system that produced them, taking different values for periodic, chaotic or stochastic signals. The nonlinear dynamics features used here were correlation dimension (D_2), correlation entropy (K_2), and Noise Level. In contrast with previous works, they were computed with a recently proposed algorithm which is based on the U-correlation integral. This algorithm provides a reliable and user-independent estimation of these invariants.[23,24]

Noise Level measures the extent at which the hypothesis for the estimation of invariants are fulfilled. In other words, when D_2 and K_2 are computed along with the Noise Level, the latter can be used as an indicator of how reliable the estimations of the other two invariants are. However, when data is acquired under controlled conditions, like the voices from the datasets used herein, Noise Level can be used as an indicator of a dominant stochastic component[23,24] assuming that its variation is caused by the proper dynamics of the signal, and not by unfulfilled hypothesis like inadequate embedding parameters. Periodic signals usually have low values for D_2, K_2, and Noise Level. In contrast, chaotic and stochastic time series might be associated with higher values. In fact, D_2 of stochastic signals is theoretically infinite. Considering this, progressively higher values can be expected for types 1, 2, and 3.

Nonlinear dynamics features like D_2, K_2 depend on the embedding parameters, ie, delay (τ) and dimension (m), used for their computation. We computed these features for several values of embedding parameters and then consider them as different descriptors during the feature selection process explained in the following section. The values of m

used were 4,6,8,10,12,14 and the values of τ used were 25,50,80,110.

Nonlinear dynamics features, NPCV, SDNPCV, CPP were computed with *MatLab 2018a* (MathWorks Natick, MA), while HNR, Shimmer and Jitter were computed using *PRAAT version 6.0.39*.[32]

Classification

An advantage of simple classification models, like linear classifiers or classification trees, is that it can be easily understood how the data is used to retrieve the corresponding class from it. This is quite important in a biomedical context, where clinicians could use this information to understand how an automatic tool for classification works. When an effective classification can be achieved using only a hyperplane in the feature space, features tend to represent better the key aspects of the classes involved. Bearing this in mind, a linear Support Vector Machine (SVM) classifier[33] was used in order to emphasize the role of the features computed and to determine if they represent the different types of voices in an objective way.

Classification performance was computed using K-fold cross-validation (with $K = 10$). A linear SVM was trained using $K-1$ folds from one dataset, and tested with the remaining fold. For cross-dataset performance estimation, the whole data from the other dataset was also used as another test set (and vice-versa). By this procedure, the generalizability of the proposed approach can be evaluated directly and thus lead to more relevant results, since this scenario is closer to a real-world situation. All data were centered in the mean and normalized in variance before training.

In order to reduce bias towards the prevalent class (type 2 voices), misclassification costs were used to penalize the classification of types 1 and 3 voices as type 2. The accuracy for each class was considered as the main performance measure as well as the overall accuracy of the model. A well trained model should have a high accuracy and also similar accuracy values for each class in order to deter the system from being biased.

RESULTS

Features selection

Nearest Neighbor Analysis (NNA)[34] and forward feature selection[35] were used to reduce the number of parameters employed and to select only those relevant to the classification. The outcome of the NNA method is a weighting vector that allows to sort the characteristics from the most relevant to the least. Following this procedure, the final combination of features used was found by forward feature selection. This method subsequently adds features one by one, starting from the most relevant according to the NNA method, while measures the accuracy of the classification model for each new combination of features. If the lastly incorporated feature increase the accuracy of classification, then it is kept. Otherwise is replaced by another feature until a better combination of traits is found. The features were added following the order previously determined by NNA.

TABLE 2.
Nearest Neighbor Analysis Outcome. Features Were Sorted From the Highest Weight to the Lowest. This Table Shows the First 20 Features and Their Corresponding Weights

NNA Weights			
Feature	Weight	Feature	Weight
CPP	1.62	Shimmer	2.24×10^{-29}
SDNPCV	1.25	$K_2(4, 25)$	2.15×10^{-31}
NPCV	1.06	$K_2(4, 50)$	1.12×10^{-31}
HNR	0.89	$D_2(6, 80)$	2.85×10^{-32}
$D_2(10, 50)$	1.13×10^{-13}	$K_2(6, 25)$	3.32×10^{-34}
Noise Level	2.94×10^{-17}	$K_2(4, 110)$	4.46×10^{-36}
$D_2(10, 25)$	2.53×10^{-20}	$K_2(4, 80)$	6.91×10^{-37}
$D_2(8, 50)$	7.36×10^{-28}	$D_2(6, 110)$	1.74×10^{-38}
$D_2(6, 25)$	1.30×10^{-28}	$D_2(4, 50)$	2.99×10^{-39}
$D_2(8, 80)$	5.02×10^{-29}	$D_2(12, 25)$	2.15×10^{-39}

Given that D_2, K_2 depend on the values of the embedding delay (τ) and embedding dimension (m), those features were noted $D_2(m, \tau)$ and $K_2(m, \tau)$ just to make explicit the parameters used for their computation. The number of nonlinear dynamics features considered was 48 (24 correlation dimension values and 24 entropy values), giving a total amount of 55 features when the other described parameters are considered as well.

Table 2 shows the first 20 weights computed by the NNA method and their corresponding features. After the forward feature selection procedure only eight features were finally selected: CPP, SDNPCV, NPCV, HNR, Noise Level, $D_2(10, 25)$, Shimmer, and $K_2(6, 25)$.

Statistical analysis

A statistical analysis was conducted on the selected features to evaluate their discrimination capabilities among all signal types. Since data distribution was not normal for any feature, a Kruskal-Wallis nonparametric test was used with a significance level of 0.005. In case the test was significant, pairwise tests were conducted between each pair of types. All tests were computed with *MatLab 2018a* (MathWorks Natick, MA). Results from this analysis can be seen in Table 3. It can be seen that almost all features take significantly different values for the three voice types ($P < 0.001$). Exceptions to this are $D_2(10, 25)$ ($P = 0.502$ for type 2 vs type 3) and $K_2(6, 25)$ ($P = 0.098$ for type 1 vs type 2) for MEEI database.

From a pattern recognition perspective, features should be uncorrelated with each other. It can be seen from the correlation matrices in Tables 4 and 5 that almost all pairs of features are not highly correlated for both datasets. This could mean that the proposed traits carry different

TABLE 3.
Statistical Analysis Outcome. Kruskal-Wallis (K-W) Was Computed Along With Pairwise Comparisons

	CPP		SDNPCV		NPCV		HNR	
	MEEI	SAAR	MEEI	SAAR	MEEI	SAAR	MEEI	SAAR
K-W	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$
1 vs 2	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$
1 vs 3	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$
2 vs 3	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$
	Noise Level		$D_2(10, 25)$		Shimmer		$K_2(6, 25)$	
	MEEI	SAAR	MEEI	SAAR	MEEI	SAAR	MEEI	SAAR
K-W	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$
1 vs 2	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P = 0.098$	$P < 0.001$
1 vs 3	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$
2 vs 3	$P < 0.001$	$P < 0.001$	$P = 0.502$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$	$P < 0.001$

TABLE 4.
Correlation Matrix of the Features For MEEI

	Correlation Matrix For MEEI Dataset							
	CPP	SD- NPCV	NPCV	HNR	Noise L.	D_2 (10,25)	Shimm.	K_2 (6,25)
CPP	1.00	−0.64	0.57	0.76	−0.55	−0.17	−0.37	−0.18
SDNPCV		1.00	−0.70	−0.78	0.63	0.19	0.45	0.23
NPCV			1.00	0.83	−0.81	−0.17	−0.74	−0.23
HNR				1.00	−0.78	−0.21	−0.56	−0.29
Noise L.					1.00	0.18	0.62	0.28
$D_2(10, 25)$						1.00	0.03	0.10
Shimm.							1.00	0.11
$K_2(6, 25)$								1.00

TABLE 5.
Correlation Matrix of the Features For SAAR

	Correlation Matrix for SAAR Dataset							
	CPP	SD- NPCV	NPCV	HNR	Noise Level	D2 (10,25)	Shimm.	K2 (6,25)
CPP	1.00	−0.65	0.48	0.78	−0.54	−0.26	−0.53	−0.29
SDNPCV		1.00	−0.53	−0.74	0.57	0.30	0.56	0.27
NPCV			1.00	0.68	−0.70	−0.19	−0.64	−0.19
HNR				1.00	−0.79	−0.35	−0.75	−0.35
Noise L.					1.00	0.26	0.70	0.29
$D_2(10, 25)$						1.00	0.25	0.11
Shimm.							1.00	0.18
$K_2(6, 25)$								1.00

information relevant for the classification. Coefficients with an absolute value above or equal to 0.75 were found for the following pairs of features. For MEEI dataset, HNR with CPP, SDNPCV, NPCV, and Noise Level (with correlation coefficients of 0.76, –0.78, 0.83, and –0.78 respectively). For SAAR dataset, HNR and CPP, Noise Level, and Shimmer (with correlation coefficients of 0.78, –0.79, and –0.75 respectively).

Classification

Tables 6 and 7 shows the performance of the classifier for intra and inter-dataset trials, when training with MEEI and SAAR respectively. It can be seen from Table 6 that the difference between the overall accuracy of the system for the intra-dataset case (87.06 %) and the inter-dataset case (86.53 %) are neglectable for MEEI. The same can be said for the SAAR dataset (Table 7), with an overall accuracy of

TABLE 6.
Confusion Matrix of the Multiclass Classifier Computed by 10-Fold Cross-Validation. For the intra-dataset trial, the training data corresponded to 9 folds from MEEI dataset while test data were from the remaining fold.*

		Test dataset: MEEI					
		Train: MEEI			Train: SAAR		
		T 1	T 2	T 3	T 1	T 2	T 3
True class	T1	90.84 (8.52)	9.15 (8.52)	0	81.95 (1.40)	18.05 (1.40)	0
	T2	11.09 (5.10)	82.93 (5.98)	5.98 (4.25)	7.94 (0.96)	87.28 (1.07)	4.78 (0.64)
	T3	0	7.76 (6.28)	92.24 (6.28)	0	8.84 (1.22)	91.16 (1.22)
Overall accuracy			87.06 (4.58)			86.53 (0.60)	

* For in the inter-dataset trial, 9 folds from SAAR dataset were used during training and the whole MEEI dataset was used as test. Values are given as *mean* (*standard deviation*)%.

83.36% for intra-dataset case and 82.71% for inter-dataset case.

For MEEI dataset (Table 6), although the accuracy of the inter-dataset and intra-dataset cases do not significantly differ, there are differences among the classification rates of types 1 and 2 for both cases. This is higher for type 1 voices than for type 2 for the intra-dataset case (90.84% and 82.93% for type 1 and 2, respectively), while the reverse is true for the inter-dataset case (81.95% and 87.28%).

By comparing Tables 6 and 7 it can be seen that performance obtained with MEEI is higher than the one obtained with SAAR.

Table 8 shows the performance of the classifier when all the signals are joined in a single dataset. The overall accuracy in this is case is 82.96%, although the percentage of correctly classified type 2 signals is 77.90%.

Noticeably, the percentage of type 3 voices correctly classified is over 90% in all trials and for all datasets. In addition, there is not confusion between type 1 and type 3 voices in any trial.

Using posterior probabilities as a measure of reliability

Since there is not a perfect system for signal typing, the possibility that the predicted type can be wrong must be considered, along with actions that can be taken to prevent misclassification, before employing an automatic software in the clinical context. Taking this into account, we propose the use of posterior probabilities of the classifier as a measure of the reliability of the automatically assigned type.

Class-wise posteriors are the probabilities that a given signal is of a certain class once all its parameters are known. These probabilities can be estimated using the linear SVM scores,[36] and the class with the highest posterior probability is the finally predicted type.

A threshold value can be computed to make the use of posterior probabilities easier. If the posterior probability surpass the threshold value, the predicted signal type can be accepted with high confidence. On the contrary, if it is too far below the threshold value, a clinician could decide to revise the automatic classification.

Using the model that was trained with both datasets, described in Table 8, the highest posterior probability of each signal from the test folds was saved. Then posterior probabilities for correctly and wrongly classified signals were considered random variables, and their probability distributions were estimated in order to determine a proper threshold. Figure 1 shows both distributions, and the computed threshold $p_{thr} = 0.7990$. It can be seen that the posterior probability distribution for correctly classified signals is

TABLE 7.
Confusion Matrix of the Multiclass Classifier Computed by 10-fold Cross-Validation. For the intra-dataset trial, the training data corresponded to ninefolds from SAAR dataset while test data were from the remaining fold.*

		Test: SAAR					
		Train: SAAR			Train: MEEI		
		T 1	T 2	T 3	T 1	T 2	T 3
True class	T1	82.78 (10.52)	17.22 (10.52)	0	84.48 (1.01)	15.52 (1.01)	0
	T2	11.82 (4.83)	80.91 (6.55)	7.27 (3.83)	14.67 (0.92)	77.69 (1.09)	7.63 (0.37)
	T3	0	8.36 (8.23)	91.64 (8.23)	0	4.95 (0.89)	95.05 (0.89)
Overall accuracy			83.36 (4.49)			82.71 (0.43)	

* For in the inter-dataset trial, ninefolds from MEEI dataset were used during training and the whole SAAR dataset was used as test. Values are given as mean (standard deviation) %

TABLE 8.
Confusion Matrix of the Multiclass Classifier Computed by 10-Fold Cross-Validation Using Both Datasets, ie, MEEI and SAAR, Combined. Values Are Given as Mean (Standard Deviation)%

Training and Test Dataset: SAAR + MEEI				
		T 1	T 2	T 3
True class	T1	86.08 (3.90)	13.92 (3.90)	0
	T2	15.03 (4.68)	77.90 (4.16)	7.07 (2.11)
	T3	0	7.57 (3.36)	92.43 (3.36)
Overall accuracy			82.96 (1.87)	

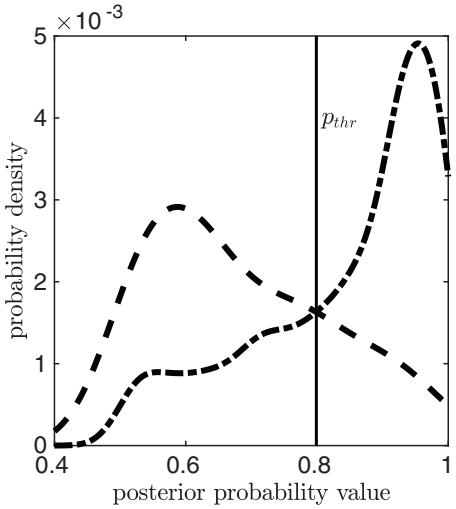

FIGURE 1. Probability densities of the posterior probability of wrongly (dashed line) and correctly (dash-dotted line) classified signals.

concentrated near 1, while the distribution for wrongly classified voices is concentrated near lower values (approximately 0.6).

The value of p_{thr} shown in Figure 1 was computed such that the probability of accepting a misclassified signal with a posterior probability above the threshold is minimized, while the same probability for correctly classified signals is maximized.

DISCUSSION

Our results show that the proposed approach could be used to successfully distinguish among the three voice signal types with state-of-the-art performance,[16] while using a linear multi-class classifier. Employing linear support vectors machines can be advantageous for the interpretation of how the predictors are used in order to classify the signals, which is a valuable property in medical contexts. Utilizing well-known parameters within the voice pathologists community could also make the here proposed method more attractive, as well as the use of posterior probabilities as a measure of the soundness of the predicted type.

In contrast to previous works,[18,19] the nonlinear dynamics features were estimated with a user-independent method[23] and the embedding parameters were chosen to maximize the performance of the classifier using an automatic feature selection method.

The novel features proposed in this article, NPCV and SDNPCV, were among the features with highest weighting factor from the Nearest Neighbor Analysis method, meaning that these features are indeed relevant for this classification problem. Moreover, statistically significant differences were found for all types of voices, for both features. These measures were proposed to describe the regularity of a signal waveform, being sensitive to changes in the overall morphology of nearly periodic signals. A PRAAT script to compute NPCV has been made available for clinicians in a public repository.[1]

The high correlation coefficients found between HNR and other features like CPP, Noise Level, NPCV, SDNPCV and Shimmer might be mainly due to the fact that all those parameters can be affected by the aperiodic component of the signal. However, most coefficients from Tables 4 and 5 are low, which might mean that the used features share little common information.

The results here reported were estimated over a larger amount of signals than all previous works,[16,18,19] and using two datasets with speakers of different languages. This is relevant because a more accurate estimate of the proposed approach's performance can be found when the system is faced with a situation that is closer to a real-word application. Although all features showed statistically significant differences for almost all signal types and datasets, like many other parameters described before,[16,18,19] results reveal that the performance of a system when trained with one corpus cannot be directly extrapolated to other datasets.

For instance, Table 6 shows that performance is lower when the system is trained with one dataset and then tested with another corpus. In addition, overall accuracy was systematically lower for the SAAR dataset than for MEEI. This leads to the conclusion that, at least for the problem of voice signal typing, assumptions on the generalizability of a system or parameter should be made cautiously when only a few signals coming from a single corpus are employed. Looking forward to their use in the clinical context, novel features and algorithms proposed for automatic and objective signal typing should be evaluated using larger datasets in the future and, when possible, different corpora.

Tables 6 –8 show that the used features can discriminate type 3 voices from the other two types with significant success, even in inter-dataset trials. This may not come as a surprise since type 3 voices correspond to highly disorganized signals that easily deviates all the parameters from the values they would take for signals with some degree of periodicity. Moreover, not a single type 3 signal was

[1] https://github.com/jmiramont/PRAAT-scripts

classified as type 1, which is desirable since type 3 voices must not be analyzed with perturbation measures under any circumstances.

In contrast, distinguishing between type 1 and type 2 signals seems to be a more challenging task, which is reflected in the percentage of confusion between those types. Further efforts might seek to study this particular problem.

Additionally, Tables 6 and 7 also suggest that there was not a major difference between both datasets in terms of the languages of the speakers (English for MEEI and German for SAAR). This could encourage the applicability of the proposed method to other languages in future works. Also, further studies should consider other vowels as well.

The use of posterior probabilities as a reliability measure of the predicted type was also evaluated. A threshold was computed so that the type assigned can be confidently accepted by the user if posterior probability surpass it. Otherwise, if the posterior probability is below this threshold, the user could decide to verify the predicted type using the traditional methods, and eventually change it. The reported threshold (Figure 1) was estimated using the model trained with all signals, since it was the more general model possible. It is a fairly conservative threshold (ie, a high value) since the classification model is not perfect.

In the future, systems with a better performance that use posterior probabilities as a reliability measure might have more liberal thresholds (ie, a lower value), since the modes of the distributions showed in Figure 1 are expected to be more separated for systems with greater accuracy. This will reduce the number of revised predictions, making a further step towards an automatic tool for signal typing.

CONCLUSION

A pattern recognition approach was proposed for automatic classification of voice signals in the three types proposed by Titze,[1] using well-known descriptors like CPP, HNR and Shimmer, in combination with nonlinear dynamics parameters and two novel features. A support vector machine was trained with a subset of features found by a feature selection strategy, so as to keep those that were more relevant to problem. Used signals came from two different corpora, one with English speakers and the other with German speakers, and intra-dataset and inter-datasets performance estimators were computed. Outcomes show that the proposed features could be used as objective measures for signal typing, in combination with a linear SVM, with a relatively high accuracy (over 82.71%), even when the system is trained and tested using different corpora. Additionally, the use of posterior probabilities was proposed as a measure of the reliability of the predicted type, to help a potential user to make an informed decision, and to bring the possibility of revise the assigned type to reduce misclassification rates. These results might pave the way to a fully automatic tool for signal typing that could be used by clinicians in the future.

ACKNOWLEDGMENTS

This work was supported by the National Scientific and Technical Research Council (CONICET) of Argentina, the National University of Entre Ríos (UNER), Instituto de Investigacion y Desarrollo en Bioingeniera y Bioinformática (IBB), and Grants PID-6171 (UNER) and PIO-1462014 0100014CO (CITER-CONICET). The authors would like to express their sincere gratitude to Dimitar Deliyski for his selfless help.

REFERENCES

1. Titze IR. Workshop on acoustic voice analysis: summary statement. *National Center for Voice and Speech*1995.
2. Behrman A, Agresti CJ, Blumstein E, et al. Microphone and electroglottographic data from dysphonic patients: type 1, 2 and 3 signals. *J Voice*. 1998;12:249–260.
3. KarneU MP, Chang A, Smith A, et al. Impact of signal type on validity of voice perturbation measures. *NCVS Status Progr Rep*. 1997;91.
4. Schoentgen J. Stochastic models of jitter. *J Acoust Soc Am*. 2001;109:1631–1650.
5. Bielamowicz S, et al. A comparison of voice analysis systems for perturbation measurement. *J Acoust Soc Am*. 1996;93.2337−2337
6. Zhang Y, Jiang JJ. Acoustic analyses of sustained and running voices from patients with laryngeal pathologies. *J Voice*. 2008;22:1–9.
7. Shaw HS, Deliyski DD. Mucosal wave: a normophonic study across visualization techniques. *J Voice*. 2008;22:23–33.
8. Choi SH, Zhang Yu, Jiang JJ, et al. Nonlinear dynamic-based analysis of severe dysphonia in patients with vocal fold scar and sulcus vocalis. *J Voice*. 2012;26:566–576.
9. Fabris C, De Colle W, Sparacino G. Voice disorders assessed by (cross-) sample entropy of electroglottogram and microphone signals. *Biomed Signal Process Control*. 2013;8:920–926.
10. Stone D, McCabe P, Palme CE, et al. Voice outcomes after transoral laser microsurgery for early glottic cancer - considering signal type and smoothed cepstral peak prominence. *J Voice*. 2015;29:370–381.
11. Freitas SV, et al. Integrating voice evaluation: correlation between acoustic and audio-perceptual measures. *J Voice*. 2015;29.390.e1−390.e7
12. Barsties B, Hoffmann U, Maryn Y. The evaluation of voice quality via signal typing in voice using narrowband spectrograms. *Laryngo-rhino-otologie*. 2016;95:105–111.
13. Sprecher AJ, Olszewski A, Jiang JJ, et al. Updating signal typing in voice: addition of type 4 signals. *J Acoust Soc Am*. 2010;127:3710–3716.
14. Mendes-Laureano JA, Moro-Velázqueza L, Gómez-García J, et al. Emulating the perceptual capabilities of a human evaluator to map the GRB scale for the assessment of voice disorders. *Eng Appl Artif Intel*. 2019;82:236–251.
15. Gómez-García JA, Moro-Velázquez L, Godino-Llorentea JI. On the design of automatic voice condition analysis systems. part i: review of concepts and an insight to the state of the art. *Biomed Signal Process Control*. 2019;51:181–199.
16. Lee JY. Parameter estimations for signal type classification of korean disordered voices. *Int J Eng Sci Technol*. 2016;7:1977–1988.
17. Zhang Y, Jiang JJ. Nonlinear dynamic analysis in signal typing of pathological human voices. *Electron Lett*. 2003;39:1021–1023.
18. Lin L, Calawerts W, Dodd K, et al. An objective parameter for quantifying the turbulent noise portion of voice signals. *J Voice*. 2016;30:664–669.
19. Calawerts WM, Lin L, Sprott JC, et al. Using rate of divergence as an objective measure to differentiate between voice signal types based on the amount of disorder in the signal. *J Voice*. 2017;31:16–23.
20. Liu B, Polce E, Jiang JJ. An objective parameter to classify voice signals based on variation in energy distribution. *J. of Voice*. 2018.

21. Liu B, Polce E, Sprott JC, et al. Applied chaos level test for validation of signal conditions underlying optimal performance of voice classification methods. *J Speech Lang Hear R*. 2018;61:1130–1139.
22. Liu B, et al. Quantification of voice type components present in human phonation using a modified diffusive chaos technique. *Ann Otol Rhinol Laryngol*. 2019.
23. Restrepo JF, Schlotthauer G. Invariant measures based on the u-correlation integral: an application to the study of human voice. *Complexity*. 2018.
24. Restrepo JF, Schlotthauer G. Automatic estimation of attractor invariants. *Nonlin Dyn*. 2018;91:1681–1696.
25. [data]. *Massachussets Eye and Ear Infirmary, Voice disorders database, version 1.03 (cd-rom)*. Lincoln Park, NJ: Kay Elemetrics Corporation; 1994.
26. [data], Barry WJ, Putzer M, Saarbrüecken voice database. 2019. Accessed: August 25 [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de/index.php4.
27. Murphy P, Akande O. Cepstrum-based harmonics-to-noise ratio measurement in voiced speech. *Nonlinear Speech Modeling and Applications*. Heidelberg: Springer Berlin; 2005:199–218.
28. Yumoto E, Gould WJ, Baer T. Harmonics-to-noise ratio as an index of the degree of hoarseness. *J Acoust Soc Am*. 1982;71:1544–1550.
29. Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality dysphonic voices and continuous speech. *J Speech Lang Hear R*. 1996;39:311–321.
30. Anand S, Kopf LM, Shrivastav R, et al. Using pitch height and pitch strength to characterize type 1, 2, and 3 voice signals. *J. of Voice*. 2019.
31. Liu B, Polce E, Jiang JJ. Application of local intrinsic dimension for acoustical analysis of voice signal components. *Ann Otol Rhinol Laryngol*. 2018;127:588–597.
32. Boersma P, Van Heuven V. Speak and unspeak with PRAAT. *Glot Int*. 2001;5:341–347.
33. Cortes C, Vapnik V. Support-vector networks. *Machine learning*. 1995;20:273–297.
34. Yang W, Wang K, Zuo W. Neighborhood component feature selection for high-dimensional data. *J Comput*. 2012;7:161–168.
35. Duda R, Hart PE, Stork DG. *Pattern Classification*. Wiley; 2012.
36. Platt J, et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Classifiers*. 1999;10:61–74.