Understanding of vocal tract pathology using speech signals analysis

Ryszard Tadeusiewicz ¹, Wieslaw Wszolek ², Andrzej Izworski ¹, Tadeusz Wszolek ².
University of Mining and Metallurgy, 30-059 Kraków, Al. Mickiewicza 30, Poland
¹ Department of Automatic
² Department of Mechanics and Vibroacoustic

Abstract

In the present work we introduce a new approach to pathological speech processing methods and recognition of various speech pathologies. We no longer attempt to recognise forms of pathological speech deformations, because it is impossible to show any compact template of a normal speech signal (as reference) and it is also impossible to show a standard form of any deformation. Almost any sample of speech, both normal and pathological, is unique — and we must introduce a new concept for the treatment of pathological speech; artificial understanding of the nature (and source) of the pathology of speech instead of traditional classification or recognition of signals. We try to implement the process for the artificial understanding of the observed forms of speech deformations with a very close look at vocal tract pathologies. Selected excerpts of research are presented concerning the application of (modified) acoustic signal processing methods and (in particular) the neural network techniques specially designed for solving the problem of "understanding" selected pathologies of vocal tract.

Keywords: speech analysis, pathological speech, vocal tract pathology, artificial intelligence

1. Introduction

Tasks related to the analysis and recognition of a pathological acoustic signal of speech, characterising selected pathological states, are exceptionally difficult. It happens that minor pathological elements (e.g. occlusion defect) strongly manifest in the speech signal, while very serious pathological changes (e.g. tumour) give only a weak and hardly readable picture of speech disturbances. In spite of the existence of multiple examples of successful automated speech recognition in the semantic (recognition of the utterance contents for e.g. voice control of machines and devices) or personal aspect (verification and identification of persons by using their speech samples), it remains very difficult to diagnose the condition and pathological changes of the voice tract using a speech signal [1]. Additionally, there is no simple way to transfer the experience related to diagnosis of the technological system, because the problems of pathological speech diagnosis are specific by the fact that for such tasks it is very difficult to find an appropriate rule for the preliminary signal analysis [2]. On the basis of the statement that for the cases of analysis of speech pathology forms and sources discussed here, the well-known methods of automated signal recognition cannot be applied. Thus, the authors propose in the present paper a completely new approach, based on the concept of automated understanding. Because of possible multiple meanings of that phrase it should be stressed that the meaning used in the presented work concerns itself with the automated understanding of the nature and character of the pathological speech signal deformations.

The exact meaning of the term understanding has no connections with the frequently discussed problem of semantic understanding of the speech signal (i.e. the contents of the pronounced sentences). As it is known the understanding in general differs from recognition by the fact that it is very strongly based on knowledge. To
clarify, the term "automated understanding" discussed here denotes such a deformed speech signal analysis, which is oriented towards revealing the sources of the observed signal forms and not towards bare analysis of these forms and diagnostic deduction based on their typology. Every person speaks in a somewhat different way, various (with respect to the contents or speed) utterances of the same person reveal various phonetic and acoustic features of his/her speech signal, and even various registrations of the same utterance recorded from the same person but in various days, can be very different. It is almost a rule, that various samples of a regular speech signal exhibit a greater variety of measurable acoustic parameters, than the measurable differences of the same parameter between these samples and the registered samples of speech, which is obviously pathological.

All this is the reason that one cannot solely rely on models of pathological speech signal recognition in a space based on the set of its features, but that in every case one should try to understand, how such a phonetic or acoustic phenomenon occurred. This means, that the diagnostic system must contain an internal model of the signal generator, based on knowledge of the speech signal and the ways of its generation - in regular and pathological conditions. It should be noticed that such a way of signal analysis closely reflects the contemporary views on the essence of the human perception of various informations from the environment. The perception concept sketched here is widely known, but the authors’ original contribution was its application as a standard for a construction of the system of automated understanding of selected voice system pathologies, treated as objects which are being diagnosed by the analysis of the acoustic speech signal, deformed by these pathologies.

The described concept includes a series of elements unquestionably difficult in practical realisation. For in the traditional way of solving diagnostic problems, the answer is frequently found more easily. However the authors’ long-time experience in the problems related to analysis, evaluation, and the classification of pathological speech signals have proved that for this task a really a new approach is required.

2. Research samples

The experimental studies of the new methods of automatic vocal tract pathologies based on speech articulation dysfunction have been carried out on persons treated for the larynx cancer (men after various types of operations). The final acoustic material has been collected from 95 persons divided into two groups:

- The reference group
- The group of patients

Both the patients and the persons from the reference group pronounced the same text (three times), which consisted of: separated polish vowels (A, U, E, I), and words containing these vowels.

The registration of the acoustic signal has been done in an anechoic chamber, where on a digital magnetic recorder the time dependencies of the acoustic pressure has been registered during the utterance of the test sample.

After registration, the signal samples were prepared in many ways, using a wide collection of signal processing methods (both in time and spectral domain), to finally obtain key elements for building algorithms realising automatic understanding of the nature of speech deformation.

3. The concept of new methodology

The widely known and simple concepts of sound pattern recognition satisfy their tasks in the routine recognition (comprehension) of the utterance contents or in verification of the speaking person.
However they do not meet expectations in the classification of various forms of speech pathology. The reason lies in the great variability and diversity of the acoustic signal of pathological speech. As a result the attempts of finding such feature space, in which a representative description and an effective discrimination between particular forms of speech pathology would be possible, encounter very serious difficulties (problems).

The key for solution of the emerging problems can be looked for in specific modelling of the human cognitive processes. The perfection of the human perception, the emulation of which is difficult, particularly for complex phenomena and transient processes (and exactly such are encountered in the context of pathological speech evaluation), results from the fact, that human brain does not work out its decisions using only a precise analysis of the received signals. The human perception always follows from the interaction of two information streams: the inner one (generated by the stored knowledge resources) and the one coming from outside (as a stream of information provided by the reception organs). Matching of these two streams, obtained by appropriate manipulation of internal model of the observed phenomenon, enables the correct perception and reliable recognition. Choosing one of many stored models of the recognised phenomena and processes, as the one to which the new sensory experience seems to fit most perfectly, the human brain accomplishes the categorisation of that experience. By performing the adaptation and modification of the model, necessary for achieving the required conformity with the sensory experience perceived, the brain indirectly measures the "distance" between the present sensory impression from its standard model. In particular the latter process seems to be interesting in the aspect related to the tasks of evaluation of the deformation level of pathological speech. The idea of using the above mentioned scheme of "cognitive resonance" as a basis for the system of automated diagnosis of pathological speech means, that the diagnostic system has to be furnished with an internal model of the signal generator. Such a model has to include the knowledge about the speech signal and the ways of its generation - both in regular and pathological conditions. The parameters of the model (based on the knowledge about the speech signal) are modified by process of the input signal analysis. In fact both information streams are being processed - the one coming from the external source to the interior of the system and the one transferred from the internal generator to the comparator accomplishing the necessary comparisons and negotiations. During that processing the actual speech signal is being analysed in a way leading to recognition of features which enable the selection of a proper "internal hypotheses generator" and such a selection of its parameters which ensures that the internal model's output is compatible with the presently perceived impression. At the same time the representation of the internal generator has to be processed too, in order to find a proper "common ground" for the required comparison of both signals. It is obvious that depending on the form and type of the deformation being studied various "grounds" will be used and the criteria used for comparison with the results of the internal generators work will also vary. For that problem it is necessary to find an intermediate solution, placed between extreme signal representations, both of which have to be considered as hardly useful. In one case the attempts of direct comparison of unprocessed pathological speech input signal are sure to fail because of the well-known effect of random (to high extent) set of phase parameters of the particular signal components, leading to the fact that speech signals which are identical in the perception aspect, can be totally different when the details of their time dependencies are considered. On the other extreme one finds the final set of perception.
parameters, or parameters controlling the signal generation carried out by the model. Such parameters (e.g. the bare formant frequencies) exhibit too small sensitivity to those signal features, which prove to be meaningful for the diagnosis. Therefore attempts have been undertaken (and are still continued) to find such parameters which will turn out to be the most sensitive with respect to the interesting aspects of the speech signal, at the same time being invariant to the highest extent with respect to the unessential features (e.g. the personal characteristics in the voice of the studied person). It seems that for the considered forms of speech pathology, related to the larynx problems, such features can be looked for in the short-term changes (shorter than the prosodic variation periods) of both value and contour of the primary (laryngeal) tone. For the other forms of speech pathology other signal parameters will have to be looked for, e.g. in the also studied problems related to the cleft palate cases, it will be probably effects accompanying the nasalizations on non-nasal speech sounds, and in the problems of dental surgery and prosthetics the details of the signal's noise component. The general concept described above has been presented in Fig.1.

4. Selected details of the described concept

The general scheme of the described method, presented above, assumes that the knowledge about the signal, expressed in the form of properly adjusted parameter values, is confronted with the external (sensory) information, coming from the signal of actual pathological speech. It is worth stressing that the proposed approach is not a simple continuation of the methods applied so far, but is oriented towards creation of a new paradigm. This is not only an effect of measuring some or other features of the speech signal and respectively better or worse recognition of the selected forms of vocal tract pathologies based on those features. This is a completely new approach, based on the attempts of penetration of the causal dependencies [relations], underlying the origins of such or another form of the studied signal deformation. Therefore in contrast to the tasks of automated recognition of pathological speech (what was the object of our group's activity for many years) the new approach is called "an automated understanding" of the pathological speech signal. The key to that method is the model of the articulation process (regular or pathological), controlled by the parameters of the observed [received] signal. Therefore in the proposed concept "understanding"
denotes the matching of the general knowledge to a specific situation. Once more the role should be stressed, played in the described concept by the generator of cognitive hypotheses and the comparator. The outputs of both the model and the system for processing of the pathological speech signal are selected signal parameters. On the basis of the measured similarity [proximity] of these parameters the choice of optimal model and the correction of the assumed model parameters are done. Therefore the hypotheses generator is constructed in a way which ensures an instant output of the signal parameters when it is required. The concept, described in general above, can be realised in a system presented in Fig.2.

Fig.2 The concept's block diagram for the parallel design.

The models 1 to n have been adjusted to the previously known cases and forms of speech pathologies. From that set of models one model is selected for further evaluation on the basis of the assumed hypothesis, resulting from the preliminary sound analysis. The outputs provided by the generator (model) within that concept are the parameters and features of the generated signal. On the basis of a multi aspect comparison of the parameters of the generator's signal and the actual pathological speech signal a verification is carried out and further specification of the working hypothesis takes place. In other words the internal model, being the key to the signal understanding, is selected and verified on the basis of features detected in the real data obtained from pathological speech.

5. Simulation model of a selected vocal tract pathology

Below a specific example will be presented of the described methodology's functioning. The complex process of acoustic speech signal generation can be presented in the form of a theoretical model mapping functions performed by particular organs. It is essential for the simulation model to enable the determination of the signal spectrum, based on the geometrical parameters of the vocal tract specific for the articulation of particular speech sounds.
The basis for presentation of the model has been taken from the works [7,8,9,10]. In the simulation model three principal modules have been distinguished:

- the source of the acoustic wave \( G \), characterised by impedance \( Z_g(j\omega) \), refers to phonotory organ
- four-terminal network, characterised by transmittance \( K(j\omega) \), refers to phonotory organ
- load impedance \( Z_{lo}(j\omega) \).

which are presented in Fig.3

![Fig.3. Model block diagram simplified of the speech organs](image)

In the present work a model of larynx generator has been assumed, considered as a source of signals of frequencies \( F_0, 2F_0, 3F_0 \) etc., where \( F_0 = \frac{1}{T_0} \), and the amplitude proportion to the up and bellow pressure glottis difference The schematic diagram of which is presented in Fig.4.

![Fig.4. Model block diagram of the larynx.](image)

The introduced notation is as follows: \( F_{sou} \)-reflects a simplified envelope of the spectral characteristic \( |A_g(j\omega)| \).

\[
F_{sou}(\omega) = \frac{1}{(\omega/\omega_0)^2}
\]

While the resistance \( R_{ag,av} \) and the source's acoustic mass \( L_{ag,av} \) are taken for respective of these elements for average value of the glottis section \( A_{gav} \).

6. Results

On this stage of research the subject is limited to the comparison of pathological signal with the model created one. As a pathological vocal tract the larynx after surgery is assumed. The reference spectrum of vowel „E” is shown in Fig.5. In comparison the vowel “E” of pathological speech is presented in Fig. 6.
In the followed figures the spectrums of vowel “E” created by model for reference and pathology are shown.

The introduced concept of signal understanding consists of introduction of quantitative factors, describing the essence of the origins of signal deformation (e.g. various pathologies of the vocal tract).

The signal of pathological speech is recorded separately of each patient and after that is converted into parameter s spaces and vector features.

Fig. 5. Spectrum of utterance of vowel E – correct pronunciation

Fig. 6. Spectrum of utterance of vowel E – pathological pronunciation

Fig. 7. Simulated spectrum of E vowel - correct
The model creates signal in the spectrum and features domains. The comparison of these two signals in those domains is carrying out. The result of the evaluation is used for elaboration of such correction of the respective model parameters, which result in the greatest similarity of both signals. The magnitude of changes of the selected model parameters is a measure of the signal deformation, and the information specifying which of the model parameters induced the signal change ensuring the greatest similarity determines the level of "understanding" of the deformation origins.

6. Conclusions

The described concept includes a series of elements unquestionably difficult in practical realisation. In conclusion it can be stated, that in the field of automated diagnosis of pathological speech it is necessary to construct special methods of automated understanding of the nature of processes leading to speech deformation, which could replace the presently employed methods of typical acoustic signal analysis and recognition, and which would be fully adapted to the specificity of the considered problem. Because of that the proposed method will have to be considerably modified, in application to various specific tasks.

References


