Automatic detection of Parkinson's disease from compressed speech recordings

Juan Rafael Orozco-Arroyave[1,2], Nicanor García[2], Jesús Francisco Vargas-Bonilla[2], and Elmar Nöth[1]

[1] Pattern Recognition Lab, Friedrich-Alexander-Universität, Erlangen, Germany
[2] Faculty of Engineering, Universidad de Antioquia, Medellín, Colombia

Abstract. The impact of speech compression in the automatic classification of speakers with Parkinson's disease (PD) and healthy controls (HC) is tested. The set of codecs considered to compress the speech recordings includes G.722, G.226, GSM-EFR, AMR-WB, SILK, and Opus. A total of 100 speakers (50 with PD and 50 HC) are asked to read a text with 36 words. The recordings are compressed from bit-rates of 705.6 kbps down to 6.6 kbps. The method addressed to discriminate between speakers with PD and HC consists on the systematic segmentation of voiced and unvoiced speech frames. Each kind of frame is characterized independently. For voiced segments noise, perturbation, and cepstral features are considered. The unvoiced segments are characterized with Bark band energies and cepstral features. According to the results the codecs evaluated in this paper do not affect significantly the accuracy of the system, indicating that the addressed methodology could be used for the telemonitoring of PD patients through Internet or through the mobile communications network.

Keywords: Parkinson's disease, speech compression, speech codec, voiced/unvoiced frames, Internet, telemonitoring, mobile communications network

1 Introduction

PD is the second most prevalent neurological disorder in the world, affecting about 2% of people older than 65 years [1]. It results as the dead of dopaminergic neurons in the substantia nigra of the mid brain [2]. People with PD develop several motor problems including bradykinesia, rigidity, postural instability, and resting tremor, among others. Non-motor deficits are also present in PD patients, including negative effects in sleep, cognition, and emotion [3]. Typically, the patients with PD develop dysarthric speech and the set of symptoms observed includes reduced loudness, monopitch, reduced stress, breathy and hoarse voice, and imprecise articulation, among others. Although about 90% of PD patients develop speech impairments, it is estimated that only from 3% to 4% of them receive speech therapy. The motor problems developed by PD patients

make difficult to attend several clinical appointments, limiting their treatment to neurological revisions mainly focused on update the dose of medicine. The research community has been interested in solving such difficulties by developing computer aided tools to assess the speech of PD patients at home. In [4] the authors present a portable device to analyze the speech of PD patients. The device provides bio-feedback of the speech volume. A signal tone is sent to the patient if the vocal intensity is below an adjustable threshold. In [5] the authors present a computer based at-home testing device (AHTD). This device is developed to assess several symptoms of PD patients such as tremor, small and large bradykinesia, speech, reaction/movement times, among others. According to their findings, the incorporation of the AHTD in larger clinical studies is feasible and it could be used to follow the progress of the disease. In [6] the author analyzes the voice of PD patients considering measures such as the fundamental frequency of the voice, energy, and the sound pressure level. The skin vibrations are also recorded using an accelerometer. Recently, in [7] the authors present a portable device to evaluate the phonatory and articulatory capability of patients with PD. The device records sustained phonations and perform several analysis including noise content, stability, periodicity, and different articulation measures such as the triangular vowel space area (tVSA), the formant centralization ratio (FCR), and the vowel articulation index (VAI). Additionally, there are several studies that present different methodologies to assess the speech signals in order to discriminate PD speakers and healthy controls [8–11].

The use of portable devices for the assessment of PD patients at home is feasible from the technical point of view; however, it could be relatively expensive either for the patients or the health system. The Information and Communication Technologies (ICT) allow to think on doing telemonitoring of PD patients using different communication tools already existing in Internet. Although, there are several aspects in such new technologies and tools that have to be studied to analyze the feasibility of using them in real scenarios. For instance, there exist different communication systems that can be used for the remote evaluation of speech e.g., the mobile communications network, the Internet, and the landline, among others. All of these technologies compress the audio signals in order to transmit them through the communication channel. The compression rates depends on the technology and on the bandwidth available in the network.

This paper explores the impact of several codecs in the performance of the methodology presented in [10] to discriminate between speakers with PD and HC. Texts read by a total of 100 speakers were recorded, 50 with PD and 50 HC. The codecs considered in this study were used to compress the speech signals at different rates depending on the application. The set of codecs includes G.722 [12] which is used in voice over IP (VoIP) land-lines, G.726 [13] which is used to compress speech signals that are multiplexed into international trunks, GSM-EFR (Global System for Mobile Communications - Enhanced Full Rate) [14] which is used in mobile networks, AMR-WB (Adaptive Multi-Rate - Wide Band) [15] which is a relatively new standard for mobile networks, SILK [16] which is

the codec used for the Skype® calls, and Opus [17] which is used in VoIP calls made trough Internet and in several applications with audio streaming.

The rest of the paper is organized as follows. Section 2 provides the details of the methodology and the experiments addressed in this study. Section 3 includes the results obtained in the experiments, and finally Section 4 provides the conclusions derived from this study.

2 Experimental setup

The methodology addressed here comprises four steps. (1) The speech recordings are compressed with six different codecs. For the sake of comparisons, the original recordings i.e., without any compression, are also considered. (2) The signals are preprocessed and the voiced/unvoiced (v/uv) speech frames are segmented. (3) voiced and unvoiced frames are characterized separately, and (4) the discrimination between speech of PD patients and HC is performed using a support vector machine (SVM). Further details of each step are provided in the next subsections.

2.1 Speech recordings

A total of 100 native Spanish speakers are considered, 50 with PD and 50 HC. All of the patients were diagnosed a neurologist expert. The age of the patients ranged from 33 to 77 (mean 61 ± 9) and the age of the healthy speakers ranged from 31 to 86 (mean 60 ± 9). All of the participants were asked to read a text with 36 words. The sampling frequency was 44.1 kHz with 16 bits of resolution. The people was recorded in a sound-proof booth, with a dynamic omnidirectional microphone and a professional audio card. Note that if these recordings would be transmitted over a network, the bit-rate would be $44.1 \times 16 = 705.5$ kbps. Further details of the database can be found in [18].

2.2 Encoding - compression

The codecs used in this study compress the speech signals in order to reduce the bit-rate and thus to make a more efficient use of the network resources e.g., bandwidth. A total of six codecs are considered. G.722 and G.726 are based on the adaptive differential pulse code modulation (ADPCM) [19] method. While GSM-FR, AMR-WB, SILK, and Opus are based on the analysis-by-synthesis concept. A brief description of each codec is provided below.

ADPCM: In this method the difference between the original signal $x(n)$ and the predicted signal $\tilde{x}(n)$ is quantized. The prediction process is based on a linear prediction (LP) filter, thus the parameters of the LP filter correspond to the model of the vocal tract. The difference between the predicted signal and the original ($d(n)$) corresponds to the excitation. The parameters of the LP filter

and the excitation signal are encoded. This procedure is summarized in Figure 1. A brief description of the G.722 and G.726 codecs is provided below.

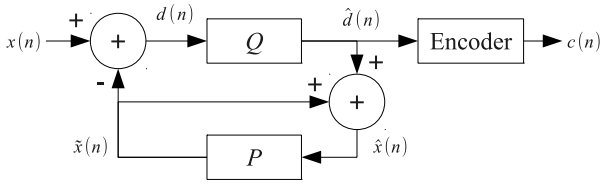

Fig. 1. General process of ADPCM. Q: quantizer, P: LP filter

G.722: This codec is defined by the International Telecommunications Union (ITU) in [12]. The spectrum of the signal is divided into two parts which are quantized independently. This codec is used in VoIP calls where high bandwidth is available e.g., land-lines and local area networks. In this case the recordings are re-sampled at 8kHz and the quantization is performed with 16 bits, thus the bit-rate is 64 kbps.

G.726: This codec is defined by the ITU [13]. It is mainly used in international trunks. It encodes the speech signal at different bit-rates. This paper only includes experiments with 16 kbps with a sampling rate of 8 kHz, which means that only 2 bits are used for the quantization of the difference $d(n)$.

Analysis-by-Synthesis: This method consists on an iterative process where the error $e(n)$ between the original signal $x(n)$ and the resulting from a synthesis model ($\hat{x}(n)$) is minimized. The parameters of the synthesis filter and the excitation signal are coded. This procedure is summarized in Figure 2. A brief description of the GSM-FR, AMR-WB, SILK, and Opus codecs is provided below.

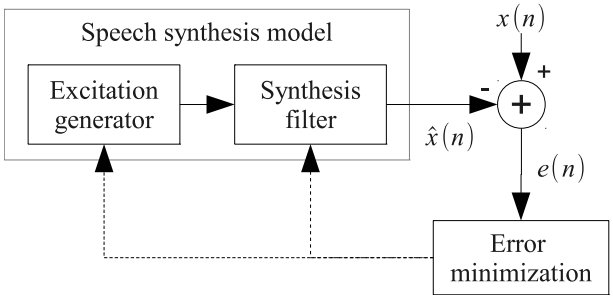

Fig. 2. General process of analysis-by-synthesis.

GSM-EFR: This codec is defined by the European Telecommunications Statandards Institute (ETSI) in [14]. It is based on the algebraic code excited

liner prediction (ACELP) encoding scheme. The bit-rate of the signals is decreased to 12.2 kbps, indicating a compression rate of 57.8 with respect to the original recordings of the database used for the experiments. This codec is widely used in the GSM mobile networks.

AMR-WB: This codec is defined by ETSI and the 3rd Generation Partnership Project (3GPP) in [15]. The standard allows to change the bit-rate over frames, however in this study only experiments with 6.6 kbps are performed. The compression rate in this case is 106.9 with respect to the bit-rate of the original recordings. This codec is being used in new implementations of GSM/UMTS mobile networks to improve the voice quality.

SILK: This codec was developed by Skype® Limited. It can encode the speech signal at variable bit-rates ranging from 6 kbps to 40 kbps [16]. For the experiments addressed in this study a sampling frequency of 24 kHz with an average bit-rate of 25 kbps is used.

Opus: This codec is defined by the Internet Engineering Task Force (IETF) in its request for comments (RFC) 6716 [17]. It is based on the SILK codec and also supports variable bit-rates which in this case range from 8 kbps to 40 kbps. This codec supports variable sampling rates. In this paper only experiments with bit-rates of 64 kbps are reported. This bit-rate is chosen in the assumption of applications with high-speed Internet connection.

2.3 Pre-processing and voiced/unvoiced segmentation

The recordings are normalized in amplitude and mean cepstral subtraction is applied to avoid possible bias introduced by the channel i.e., microphone and sound card. The segmentation of voiced and unvoiced frames is performed in Praat [20]. Voiced and unvoiced segments are grouped separately. Each frame is windowed using Hamming windows with 40 ms length and 20 ms time shift. Frames shorter than 40 ms were excluded as well as pauses longer than 270 ms.

2.4 Characterization

The *Voiced frames* are characterized with 12 MFCC along with their first and second derivatives (Δ and $\Delta\Delta$). Perturbation measures such as absolute and relative values of jitter and shimmer, and the variability of F_0 are also included. Additionally, four noise measures are considered: Harmonic-to-Noise Ratio (HNR), Glottal-to-Noise Excitation Ratio (GNE), Noise-to-Harmonic Ratio (NHR), and Normalized Noise Energy (NNE). The *Unvoiced features* are characterized with 12 MFCC, Δ, and $\Delta\Delta$. The energy content of the unvoiced frames is measured over 25 band scaled according to the Bark scale [21]. The mean value, standard deviation, kurtosis, and skewness are calculated from each feature vector.

2.5 Classification

A support vector machine (SVM) with soft margin is used to discriminate between PD and healthy speakers. The margin parameter C and the bandwidth of

the Gaussian kernel γ are optimized through a grid-search with $10^{-3} < C < 10^4$ and $10^{-1} < \gamma < 10^3$. The selection criterion was based on the accuracy obtained in the test set. The SVM is trained following a $10-$fold cross-validation strategy. All of the folds were formed randomly but assuring the balance in age, gender, and the speaker independence.

3 Results

The results are presented in terms of the accuracy obtained in the classification process. The standard deviation measured among the 10 folds in the validation process is also indicated. Table 1 includes the results with the unvoiced frames. Note that most codecs do not affect significantly the accuracy of the classifier. Only the results on GSM-EFR and Opus are slightly reduced.

Codec	SR	BR	MFCC+Δ+$\Delta\Delta$	BBE	All
Original	44.1	705.6	97 ± 7	95 ± 7	97 ± 5
G.722	16	64	97 ± 5	95 ± 7	99 ± 3
G.726	8	16	97 ± 10	94 ± 7	95 ± 10
GSM-EFR	8	12.2	94 ± 7	93 ± 10	96 ± 5
AMR-WB	16	6.6	96 ± 8	98 ± 6	95 ± 9
SILK	24	25	98 ± 4	95 ± 5	96 ± 7
Opus	variable	64	93 ± 10	95 ± 10	94 ± 11

Table 1. Classification results obtained with unvoiced frames (values in %). SR: sampling rate [kHz], BR: bit-rate [kbps], BBE: Bark band scales, All: merging all features.

The results in Table 2 show that most of the codecs do not affect significantly the accuracy of the classifier using measures extracted from the voiced segments. However, when the G.726 or SILK codecs are used, the accuracies of the perturbation measures increase with respect to those obtained with the original recordings. It seems like the modifications of the speech spectrum performed by these two codecs are affecting the frequencies above 500 Hz, but not modifying the frequencies around the fundamental frequency, which is the basis to estimate the perturbation features.

Codec	SR	BR	MFCC+Δ+$\Delta\Delta$	Noise	Perturbation	All
Original	44.1	705.6	86 ± 8	77 ± 12	76 ± 8	84 ± 11
G.722	16	64	86 ± 11	79 ± 11	77 ± 8	87 ± 8
G.726	8	16	83 ± 12	74 ± 8	80 ± 15	81 ± 11
GSM-EFR	8	12.2	86 ± 8	80 ± 9	82 ± 9	88 ± 6
AMR-WB	16	6.6	81 ± 7	76 ± 10	75 ± 13	79 ± 13
SILK	24	25	84 ± 8	70 ± 13	83 ± 13	82 ± 11
Opus	variable	64	88 ± 6	76 ± 7	77 ± 7	86 ± 7

Table 2. Classification results obtained with Voiced frames (values in %). SR: sampling rate [kHz], BR: bit-rate [kbps], All: merging all features.

The results are summarized in Figure 3. Note that the highest accuracies are obtained with the unvoiced features in all of the cases. It seems that there is almost no negative impact of the codification methods in the performance of the system.

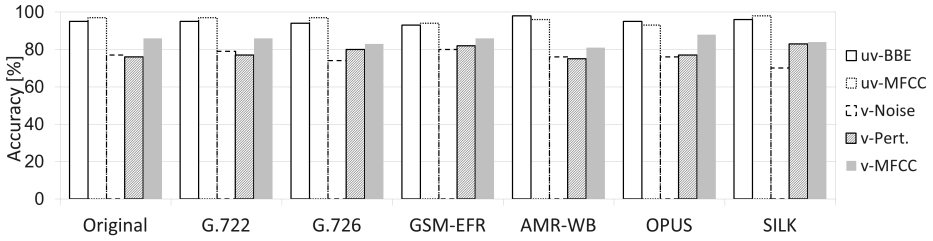

Fig. 3. Accuracy for each codec with each set of characteristics.

4 Conclusion

Speech recordings of 50 patients with PD and 50 HC are compressed considering six speech codecs widely used in different commercial applications through Internet or through the mobile network. The impact of such codecs in the automatic discrimination of speakers with PD and HC is evaluated in this paper. According to the results, the impact of the audio-compression in the accuracy of the system is minimal. Although the results indicate that the methodology addressed here could be used for telemonitoring PD patients through Internet or the mobile communications network, it is worthy to note that we did not consider the effects introduced by the communications channel, i.e., scenarios with loss of packets during the communication are not considered. Further experiments with recordings captured through Internet or through the mobile network are required to obtain more conclusive results.

References

1. M. C. de Rijk, "Prevalence of Parkinson's disease in Europe: A collaborative study of population-based cohorts," *Neurology*, vol. 54, pp. 21–23, 2000.
2. O. Hornykiewicz, "Biochemical aspects of Parkinson's disease," *Neurology*, vol. 51, no. 2, pp. S2–S9, 1998.
3. J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients," *Journal of Speech and Hearing Disorders*, vol. 43, pp. 47–57, 1978.
4. R. Rubow and e. Swift, "A microcomputer-based wearable biofeedback device to improve transfer of treatment in Parkinsonian dysarthria," *Journal of Speech and Hearing Disorders*, vol. 50, no. 2, pp. 178–185, 1985.
5. C. G. Goetz and et al., "Testing objective measures of motor impairment in early Parkinson's disease: feasibility study of an at-home testing device," *Movement Disorders*, vol. 24, no. 4, pp. 551–556, 2009.
6. M. Wirebrand, "Real-time monitoring of voice characteristics using accelerometer and microphone measurements," Master's thesis, Linkping University, Linkping, Sweden., 2011.
7. J. C. Vásquez-Correa, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth, "New computer aided device for real time analysis of speech of people with Parkinson's disease," *Fac. Ing. Univ. Antioquia*, vol. 1, no. 72, pp. 87–103, 2014.
8. J. Rusz, R. Cmejla, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Acoustic assessment of voice and speech disorders in Parkinson's disease through quick vocal test," *Movement Disorders*, vol. 26, no. 10, pp. 1951–1952, 2011.
9. T. Bocklet, S. Steidl, E. Nöth, and S. Skodda, "Automatic evaluation of Parkinson's speech - acoustic, prosodic and voice related cues," in *Proceedings of the 14th INTERSPEECH*, 2013, pp. 1149–1153.
10. J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Automatic detection of Parkinson's disease from words uttered in three different languages," in *Proceedings of the 15th INTERSPEECH*, 2014, pp. 1473–1577.
11. J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth, "Spectral and cepstral analyses for Parkinson's disease detection in Spanish vowels and words," *Expert Systems*, pp. 1–10, 2015, to appear.
12. International Telecommunication Union (ITU), *7 kHz audio-coding within 64 kbit/s. Recommendation ITU-T G.722*, Std., 2012.
13. ——, *40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM). Recommendation ITU-T G.726*, Std., 1990.
14. *Digital cellular telecommunications (Phase 2+); Enhanced Full Rate (EFR) speech transcoding; (GSM 06.60 version 8.0.1 Release 1999)*, European Telecommunications Standards Institute (ETSI) Std., November 2000.
15. *Adaptive Multi-Rate Wideband (AMR-WB) speech Codec; Transcoding functions (3GPP TS 26.190 version 12.0.0 Release 12)*, 3rd Generation Partnership Project (3GPP) Std., October 2014.
16. Internet Engineering Task Force (IETF), *SILK Speech Codec*, Std., 2010.
17. ——, *Definition of the Opus Audio Codec. RFC 6716*, Std., 2012.

18. J. Orozco-Arroyave, J. Arias-Londoño, J. Vargas-Bonilla, M. González-Rátiva, and E. Nöth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Proceedings of the 9th Language Resources and Evaluation Conference (LREC)*, 2014, pp. 342–347.
19. L. R. Rabiner and R. W. Schafer, *Introduction to digital speech processing*, 4th ed. Hanover, MA: now Publishers Inc., 2007, vol. 1, no. 1-2.
20. P. Boersma and D. Weenink, "PRAAT, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.
21. E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *Journal of Acoustical Society of America*, vol. 68, no. 5, pp. 1523–1525, 1980.