International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

http://ijpsat.ijsht-journals.org
Vol. 10 No. 1 August 2018, pp. 163-169

Pitch Marking Study towards High Quality in Concatenative Based Speech Synthesis

Margarita Selene Salazar Ávila, José Antonio Trejo Carrillo
Centro de Investigación para la Administración Educativa
Independencia 1100, Centro, 78000 San Luis, S.L.P.
Av. Cerro del Padre 907, Zacatecas, Zac.

Fabián Navarrete Rocha, Naram Isaí Hernández Belmontes, Jesús Rodríguez Zamarrón,
Javier Saldívar Pérez, Ana luisa Hernández Gutiérrez
Centro de Investigación e Innovación Automotriz de México (CIIAM)
Ramón López Velarde 801, Zacatecas Centro, 98000 Zacatecas, Zac.

César Gamboa Rosales
Universidad Aeronáutica en Querétaro
Carr. Querétaro-Tequisquiapan, Coyote, Qro.

Lorena Raquel Casanova Luna
Universidad de Tolosa
Jardín de la Madre 202, Zacatecas Centro, 98000 Zacatecas, Zac.

Abubeker Gamboa Rosales
Universidad Tecnológica del Estado de Zacatecas (UTZAC)
Km 5, Carretera Zacatecas, 98601 Cuauhtémoc, Zac.

Claudia Sifuentes Gallardo
Centro de Investigaciones en Óptica (CIO)
Lomas del Bosque 115, Lomas del Campestre, 37150 León, Gto.

***Abstract*—Generally, concatenative speech synthesis systems provide a considerable synthesis quality since the criteria for unit selection methods have been optimized. However, the level of synthesis quality still depends on the adequate concatenation of speech units. An adequate concatenation of speech units has as precondition that concatenation mismatches as phase mismatch, phase mismatch and discontinuity of spectral envelope must not appear in the synthesized speech signal. Therefore, avoiding phase mismatches leads to a high speech synthesis quality and the way to avoid phase mismatches is achieved by an appropriated pitch marking algorithm. Therefore, a pitch marking study was carried out through a evaluat-ing the available pitch marking algorithms. So, a speech database was pitch marked many times using the different pitch mark algorithms. Therewith, several sentences were synthesized applying the different pitch makings of the speech database. A mean Opinion Score (MOS) listening test was carried out for the evaluation of the synthesized speech sentences regarding mismatch human perception. The best pitch mark algorithm was selected according its observed effect in the quality of the speech synthesis.**

Keywords*— *Speech Synthesis, Unit Selection, Concatenation Mismatch

Corresponding Author: Fabián Navarrete Rocha

I. Introduction

In the unit selection of concatenative based speech synthesis systems must be consider the speech signal analysis in timeand frequency domain in order avoiding concatenation distortions as phase mismatch, pitch mismatch and spectral envelope discontinuities parameter. Specially, speech distortions known as concatenation mismatches can appear in asynthesized speech signal, when the time domain pitch synchronous overlap-add (TDPSOLA) algorithm is performed within quasi-stationary speech units extracted from different words and their spectral properties at their concatenation boundaries are contextually and phonologically different. Because, avoiding mismatches to synthesize an input text can be obtained a synthesized speech signal with high quality. The pitch marking has taken an important role by concatenative speech synthesis in order to avoid pitch mismatches.

Therefore, it is important to find out which pitch mark algorithm has the best performance in the concatenative based speech synthesis to avoid phase mismatch-es in a synthesized speech signal. Concatenate based speech synthesis bases its functionality on the algorithm PSOLA using the speech period marks of database to manipulate the speech units in its pitch and timing. For this reason, the pitch period must be marked accurately and consistently for voice segments in the speech data base, in such way that the concatenated speech units exhibit smooth transitions at their concatenation borders [1].

So, approaches as maximum or minimum peak in speech signal, zero crossing and glottis closure or maximum slope at a given electroglottograph signal are used to mark the pitch periods of a given speech signal.Normally, the adequate position of the pitch mark is a complicated issue, because it correlates the Electroglottograph (EGG) and speech signals and determines if pitch mismatches can appear or not. There are many pitch marking methods as [2], [3], [4], [5] and [6], which use the previously mentioned different signal processing algorithms. This study carries out a pitch marking test towards getting high quality speech synthesis, in order to find the most adequate pitch marking algorithm for the concatenative based speech syn-thesis. The analysis and selection of the best pitch marking algorithm is meaningfully for speech synthesis quality, be-cause every pitch marking method improves or affects the quality of the synthesized speech signal on its own way. In Section 2 are described in detail the main causes of concat-enation mismatches. Section 3 shows the main features of the pitch marking algorithms to be evaluated. Section4 shows listening test for the pitch marking evaluation. Finally, the results and conclusions are shown in Sections 5 and 6, respectively.

II. Concatenation Mismatches

The analysis of the main causes that lead to concatenation mismatches is essential to prevent its occurrence in the synthesized speech signal. Controlling the causes that provoke the appearance of concatenation mismatches lead to achieve an optimal quality of speech synthesis. Therefore, concatenation mismatches are described in detail below.

A. Phase Mismatch

The phase mismatch appears in the synthesized speech signal using TD-PSOLA, because the respective windowing ofthe speech units are not in the same positions within the periods [1].The phase mismatches are possible even if the speech units are contiguous in the speech database. For this reason, the pitch mark method must be set at the time of the glottis closure in the center of each window [7]. This task involves the following difficulties:

- Positioning windows need specialized speech scientists.
- When the positioning of the windows is automatically performs. It is not very precise.
- Positioning windows is very time consuming, if this is not done automatically.

By the above a phase mismatch through concatenating two

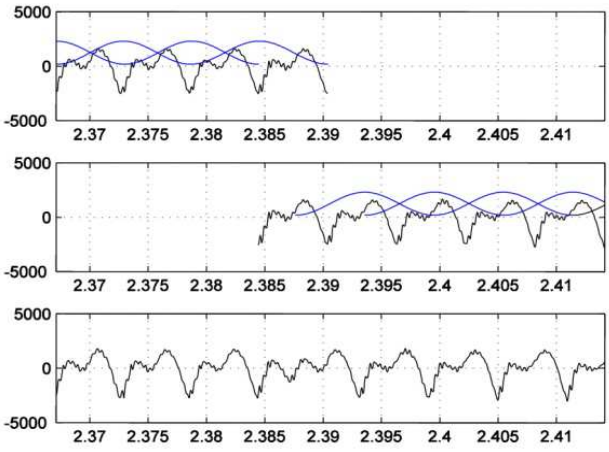

Fig. 1. Phase mismatch

Speech units is shown in the following Fig. 1.

In the Fig. 1 are shown two (left and right) speech wave-forms, which are contiguously in the speech database. Therefore, the synthesis of both speech units applying

TDPSOLA should not produce any concatenation mismatch. However, the overlap-add (OLA) windowing is not centered at the same relative positions within their periods. This has as consequence that a phase mismatch appears by concatenation both speech units as it is shown in the bottom of the Fig. 1. Because the adequate positioning of the windows is still a discussion issue, phase mismatches cannot be completely avoid, but there are already algorithms that already do a good performance [8], [9].

B. Pitch Mismatch

In the TTS voice database - system, there are speech units that show identical spectral envelopes and windows in a relatively similar position. However, these speech units are not always overlap-added, because their periods are very different. Therefore, when recording the speech database it is important for the speaker to read the text corpus with a constant pitch in order to avoid pitch mismatch between the speech segments [1].Below in Fig. 2, a pitch mismatch byconcatenating of two speech units of a given speech data is shown.

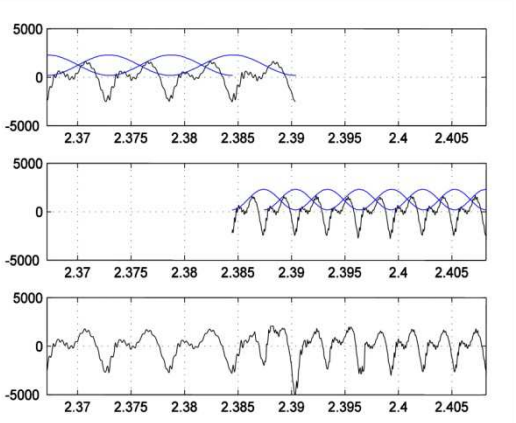

Fig. 2. Pitch Mismatch

The Fig. 2 shows left and right speech units, which do not have spectral discontinuities at their concatenation point and OLA windows were positioned in a consistent manner. However, the speech segments were spoken with different tone (pitch). This produces a pitch mismatch and degradates the synthesized signal speech quality as is shown in the bottom of the Figure.

C. Spectral Envelope Discontinuities

Discontinuities of the spectral envelope are one of the main qualityspeech signal degraders in concatenative speech synthesis. The reason is the phonetic variability and the coarticulation of the speaker reverberation that occur in every spoken language. In Fig. 3, the discontinuities of the spectral envelope are shown in detail [10], [11].

The pitch in Fig. 3 is constant and the windows were positioned in a consistent manner. The spectral discontinuity is concentrated in oneperiod. The discontinuities of the spec-tral envelope can be partially avoided by the energy normalization of the language elements.

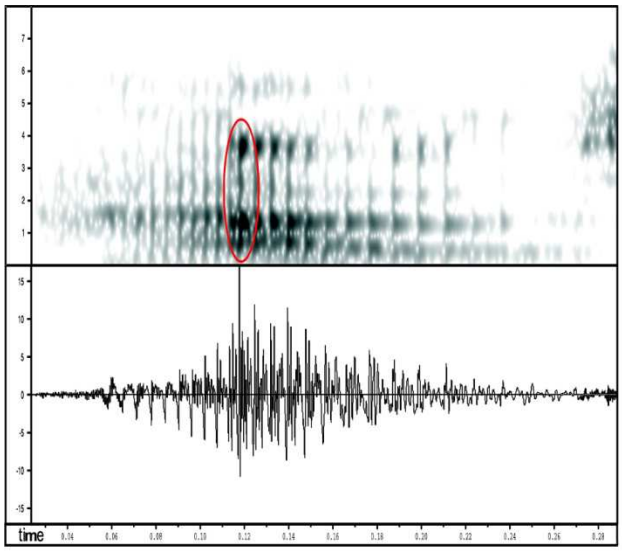

Fig. 3. Discontinuities of the Spectral Envelope.

So far, the concatenation mismatches as phase and pitch mismatch and discontinuities of the spectral envelope have been analyzed in detail in order to see their impact in the quality of the speech synthesis. Therewith, in the following Section 3 are described the pitch marking algorithms, which should contribute to avoid the phase mismatch through anadequate positioning of period windows used by PSOLA in the concatenative based speech synthesis [12], [13].

III. Description Of Pitch Mark Algorithms

Pitch marking has the highest significance in signal processing regarding PSOLA. In speech synthesis, pitch marking robust estimation is particularly important for the modification of fundamental frequency F0 and the timing of synthesized speech signal. Therewith, prosody modification can be achieved to improve the expressivity of the concatenative based speech synthesis [14].

In addition, the selection of an appropriate position of the pitch marks is determinate differently for each pitch marking algorithm, which take as reference the spectral features of the speech signal or the EGG signal. Normally, the pitch mark position is set at the beginning of glottal closure known as Glottal Closure Instant (GCI) [15], [16], [17].

A. Robust Marking Of Pitch Periods

The GCI is marked in the speech signal at the beginning of glottis closure and corrected by a nearly constant time delay, which coincides with the most negative peak of the speech signal. The robust marking algorithm has been designed to give a pitch contour reference of speech signals for a comparative evaluation of pitch marking algorithms [2]. For the purpose of determining the pitch general error, analysis and accuracy measurement have found that the accuracy of the algorithm must be more than 0.5% in order to detect inaudible fine measurement errors. The algorithm uses the out-put signal of the laryngograph that measures the laryngeal vibrations based on the change of the electrical impedance. The turning point during the rapid rise of laryngorgrammes is used as an estimate of the timing of the glottic closure. Using a local interpolation, the size of the measurement error must be less than 0.5 %.

B. Pitch Marking By Festival's Speech Tools

If no electroglottograph signal of a voice recording is available, the pitch marks are extracted alternatively by other signal processing methods [3], [4], [5]. The period marks are determined using an estimation of the fundamental frequency (F0) contour of the given speech signal. These methods performbest using clean laboratory recorded speech signals. However, although these methods can never be as well as the other methods which extract the pitch marks from the EGG signal, good results are also achieved [18], [19].

Additionally, the mentioned methods require a higher computational effort and are time consuming by the pitch marking a given speech signal, because their estimation requires relatively high-order filters. Festival speech tools include a pitch marking program, which processes the EGG signals to a set pitch marks in the given speech signal. However, this pitch marking tool is still not a fully automatic process and requires the monitoring of results and decisions to modify individual parameters, which could improve the result. The parameters and their definition for the pitch marking tool from Festival program are explained in detail by [3].

C. Pitch Marking By Praat

Praat is an open software tool for speech labeling and signal processing as well as for various acoustic analyzes such as spectral, formant, pitch marking, intensity, timing and prosodic manipulation of speech signals [4]. This language processing software allows also the user to implement articulatory speech synthesis, optimality-theory grammars, image processing, neural network simulations and discriminative statistics. Praat software is mostly used by phoneticians and speech scientists, because it is script programming based and easily to learn being a strong language processing tool in the speech research community [20].

D. Pitch marking by Snack

Snack signal processing software has been developed to be used with a script programming [5]. By applying the Snack software in speech signal processing, it is possible to create powerful multi-platform audio applications with just a few lines of code. The combination of Snack software and script programming helps to perform language processing tools and applications with minimal effort. Therefore, it is used for many projects to estimate the pitch marking of a given speech database [9], [21].

E. Pitch Marking By Hybrid Electroglottograph And Speech Signal Based Algorithm

The pitch marking algorithm proposed by [6] improves the accuracy of the automatic pitch marking algorithms using a hybrid method. The hybrid method combines the ad-vantages of electroglottograph and voice signals, extracting the period marks of the EGG signal and estimating the pitch marks of the speech signal using the F0 contour of the speech signal. Afterwards, a local interpolation between EGG extracted pitch marks and F0 contour estimated pitch marks is applied determining the hybrid pitch marks. The pitch marking method evaluation is described in [6], where the evaluation results show that the proposed hybrid method performs better than the algorithms based only on the electroglottographor voice signal regarding the signal quality of the concate-native based speech synthesis [22], [23].

IV. Pitch Marking Study

The English "TC-STAR" speech database was used for the pitch marking study. The speech database was designed and implemented with very high quality speech signal processing criteria [10]. The recordings voice quality has a precision of 16 Bits, sampling rate of 96 kHz, SNR > 40 dB and a bandwidth from 40 Hz to 20,000 Hz. The database recordings have duration of 10 hours. TC-STAR database contains 90,000 words, which are contained in 5558 sentences.

This sentences amount is distributed into sub-corpora of transcribed speech, written text, constructed phrases and expressive speech. The speech database labeling was carried automatically out for speech specialists labeled the 70% of the sentences of the corpus and rest of the corpus sentences (30%) manually [23].

A. Speech Database Pitch Marking

The TC-STAR speech database was pitch marked by the five previously mentioned algorithms. Therefore, five different pitch-making sets were estimated for the same database and every pitch marking set was label for its performance evaluation in the speech quality synthesis. The assigned labels for the pitch marking algorithms were GCIDA for the Robust pitch marking , FESTIVAL for the pitch marking tool from Festival TTS, PRAAT for the pitch marking algorithm used by the praat language processing software, SNACK for the scrip programming based speech signal processing software and HYBRID for the algorithm based on the electroglottogra-phand speech signal.

B. Concatenative Based Speech Synthesis

Once, the speech database was pitch marked for the five different methods. 100 text input sentences were random picked up from a text corpus to carry out this study [11]. The Dress TTS system was utilized to synthesize five times the one hundred sentences using the five different pitch marking sets [12]. Dress TTS is a concatenative based speech synthesis, which contains a rule-based text processing normalization, neuronal networks based speech prosody estimator, an optimal Bayes-based unit selection algorithm [13] and concatenates the selected speech units using the TDPSOLA algorithm.

Therewith, 500 synthesized sentences were obtained for later use by listening test. The TC-STAR speech database was pitch different methods. 100 text input sentences were random picked up from a text corpus to carry out this study [11]. The Dress TTS system was utilized to synthesize five times the one hundred sentences using the five different pitch marking sets [12]. Dress TTS is a concatenative based speech synthesis, which contains a rule-based text processing normalization, neuronal networks based speech prosody estimator, an optimal Bayesbased unit selection algorithm [13] and concatenates the selected speech units using the TDPSOLA algorithm. Therewith, 500 synthesized sentences were obtained for later use by listening test [24].

C. Listening Test

There were 100 synthesized sentences per pitch marking method. For the listening test, 30 synthesized sentences were selected randomly per set. So, every pitch marking method was represented for 30 selected synthesized sen-tences. Afterwards, the 150 synthesized sentences were played randomly for 25 listeners, and they were not allowed to listen more than two times every synthesized sentence. For the evaluation, listeners were asked qualify the intelligi-bility, naturalness, expressivity and general speech signal quality of the synthesized sentences in a scale from 1 (“poor”) to 5 (“excellent”) rates regarding a Mean Opinion Scale [14], [25].

V. Evaluation And Results

After the listening test, the listener MOS values war used to find out pitch marking algorithm delivers better quality speech synthesis by avoiding concatenation mismatches. Listener evaluation MOS values of the five pitch marking methods are summarized in the following Fig. 4.

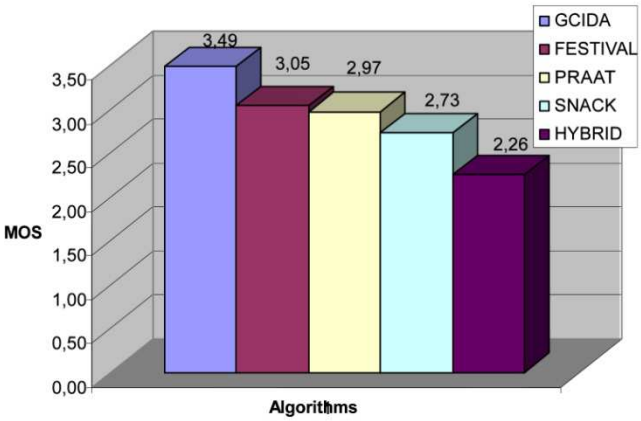

Fig. 4. Pitchmarking Algorithm Mean Scores.

The Fig. 4 shows the GCIDA-pitch marking algorithm as the best performing algorithm for quality of concatenative based speech synthesis with a total MOS score of 3.49. In second place is the pitch marking algorithm from Festival [3], which received a MOS score of 3.05. Following, the Praat and Snack pitch marking algorithms on the third and fourth place with MOS values of 2.97 and 2.73, respectively. Finally, the last place was obtained by the hybrid Pitchmarking algorithm with a MOS value of 2.26. Thereafter, the Kolmogorov-Smirnov test was applied for testing the normal distribution of the data. The influence of the different pitch marking methods was initially checked by a factorial ANOVA analysis.

Additionally, the mean score opinion results shown in the Fig. 4 were also confirmed by applying a test, whether significant mean differences between the individual methods exist. The following Table 1 illustrates the results obtained.

Table I. Significant Mean Differences (t-test)

Algorithms	Mean Differences	Significant
	3.496 - 3.046	**0.015**
GCIDA-PRAAT	**3.496 - 2.973**	**0.046**
GCIDA-SNACK	**3.496 - 2.733**	**0.010**
GCIDA-HYBRID	**3.496 - 2.260**	**0.010**
ESTIVAL-PRAA	3.046 - 2.973	0.757
FESTIVAL-SNACK	**3.046 - 2.733**	**0.044**
FESTIVAL-	**3.046 - 2.260**	**0.005**

PRAAT-SNACK	2.973 - 2.733	0.308
PRAAT-HYBRID	2.9733 - 2.260	0.027
SNACK-HYBRID	2.733 - 2.260	0.054

Table I shows that the mean differences of GCIDA over all other algorithms are significant. In addition, the mean value of FESTIVAL is significantly higher than that of Snack and HYBRID as well as the average of the mean scores of opinion of Praat is higher significantly than that of HYBRID.

Finally, the MOS values were displayed in a stem-sheet dia-gram to analyze better the relative frequency of the MOS values. In the Fig. 5 a stem-sheet diagram for the five differ-ent Pitch Marking algorithms is shown in detail.

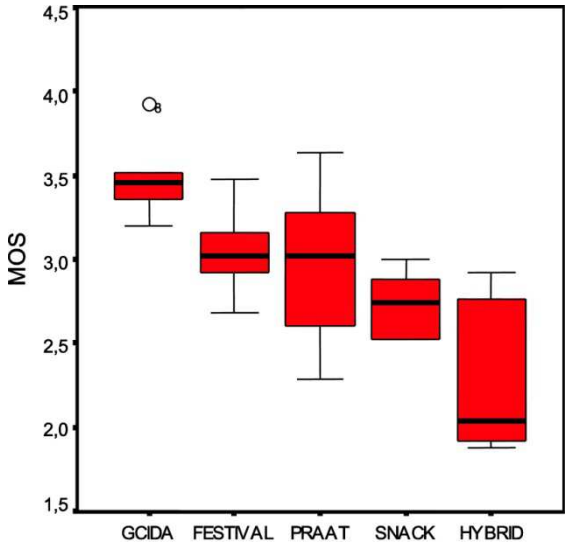

Fig. 5. Stem-sheet digram of the MOS values.

The analysis of the relative frequency of Pitchmarking algo-rithms in Figure 5 by a stem-sheet diagram shows that the GCIDA-pitch mark [2] also has the best resistance in the MOS score.

VI. Conclusions

The concatenations mismatches are the main factor for degradation the speech quality of the concatenative based speech synthesis. Therefore, this study had as main goal determine the best pitch marking algorithm that works best in the concatenative based speech synthesis for getting high speech quality and reduces the appear principal of phase mismatches. During the analysis of the results of the listen-ing test, the pitch marking method proposed by [2] has obtained the best average of the mean opinion score. Thus, this method is most suitable in order to reduce the appearance of phase mismatch caused by concatenation mismatches. Additionally, the pitch marking Festival can be also determined as a good tool for pitch marking a speech database with the advantage that festival is open-source software.

References

[1] Dutoit, T., “An introduction to Text-To-Speech Synthesis”, Kluwer Academic Publishers, 1996, 326 pp.

[2] Engel, T., “RobusteMarkierung von Grundfrequenzperioden”, Diplomarbeit, TechnischeUniversitt Dresden, 2003.

[3] Louw, A., “A short guide to pitch-marking in the Festival speech synthesis system and recommendations for improvements”. January 2004. http://www.llsti.org/pubs/Pitch marking.pdf.

[4] Boersma, P. and Weenink, D.,“Praat: doing phonetics by computer (version 4.4.19)”. http://www.praat.org, 2001. As of June 30, 2006.

[5] Sjlander, K. And Beskow J., “WaveSurfer - an open source speech tool.”, In Proc. of ICSLP, Beijing, Oct 16-20, 2000, Vol 4, pp. 464-467.

[6] Hussein, H. and Jokisch, O., “Hybrid electroglottograph and speech signal based algorithm for pitch marking”, In INTERSPEECH- 2007, pp. 1653-1656.

[7] Upadhyay, N. & Rosales, H.G. Int J Speech Technol (2016) 19: 869. https://doi.org/10.1007/s10772-016-9370-4

[8] Galván-Tejada C, López-Monteagudo F, Alonso-González O, Galván-Tejada J, Celaya-Padilla J, Gamboa-Rosales H, Magallanes-Quintanar R, Zanella-Calzada L (2018) A Generalized Model for Indoor Location Estimation Using Environmental Sound from Human Activity Recognition. ISPRS International Journal of Geo-Information 7(3), 81. doi:10.3390/ijgi7030081.

[9] Upadhyay, N. & Rosales, H.G. Natl. Acad. Sci. Lett. (2018) 41: 15. https://doi.org/10.1007/s40009-017-0597-7

[10] Nematollahi, M.A., Vorakulpipat, C., Gamboa-Rosales, H. et al. Proc. Natl. Acad. Sci., India, Sect. A Phys. Sci. (2017) 87: 433. https://doi.org/10.1007/s40010-017-0371-8

[11] Nematollahi, M.A., Gamboa-Rosales, H., Martinez-Ruiz, F.J. et al. Multimed Tools Appl (2017) 76: 7251. https://doi.org/10.1007/s11042-016-3350-1

[12] Luna-García, H., Mendoza-González, R., Gamboa-Rosales, H., Celaya-Padilla, J., Galván-Tejada, C., López-Monteagudo, F., Collazos-, C., Mendoza-González, A..(2018). MENTAL MODELS ASSOCIATED TO VOICE USER INTERFACES FOR INFOTAINMENT SYSTEMS. DYNA, 93(3). 245. DOI: http://dx.doi.org/10.6036/8766

[13]García-Hernández A, Galván-Tejada C, Galván-Tejada J, Celaya-Padilla J, Gamboa-Rosales H, Velasco-Elizondo P,

Cárdenas-Vargas R (2017) A SimilarityAnalysis of Audio Signal to Develop a Human ActivityRecognitionUsingSimilarity Networks. Sensors 17(11), 2688. doi:10.3390/s17112688.

[14] Hossmar, Maria. Master Thesis: "Emotional Prosody for Speech Synthesis". Dresden University of Technology, 2009.

[15] Kotnik, B., "Determination of characteristic points inside the glottal cycle", Technical report, University of Maribor, Faculty of Electrical. Engineering and Computer Science, 2005.

[16] Mohammad Ali Nematollahi, ChaleeVorakulpipat, and Hamurabi Gamboa Rosales, "Optimization of a Blind Speech Watermarking Technique against Amplitude Scaling," Security and Communication Networks, vol. 2017, Article ID 5454768, 13 pages, 2017. https://doi.org/10.1155/2017/5454768.

[17] Mohammad Ali Nematollahi, ChaleeVorakulpipat, and Hamurabi Gamboa Rosales, "Semifragile Speech Watermarking Based on Least Significant Bit Replacement of Line Spectral Frequencies," Mathematical Problems in Engineering, vol. 2017, Article ID 3597695, 9 pages, 2017. https://doi.org/10.1155/2017/3597695.

[18] Nematollahi, M.A., Al-Haddad, S.A.R., Doraisamy, S. et al. Natl. Acad. Sci. Lett. (2016) 39: 197. https://doi.org/10.1007/s40009-016-0430-8

[19] Carlos E. Galván-Tejada, Jorge I. Galván-Tejada, José M. Celaya-Padilla, et al., "An Analysis of Audio Features to Develop a Human Activity Recognition Model Using Genetic Algorithms, Random Forests, and Neural Networks," Mobile Information Systems, vol. 2016, Article ID 1784101, 10 pages, 2016. https://doi.org/10.1155/2016/1784101.

[20] Luna-Garcia, H., Mendoza-Gonzalez, R., Gamboa-Rosales, H., Celaya-Padilla, J., Galvan-Tejada, C., Lopez-Monteagudo, F., Collazos-, C., Mendoza-Gonzalez, A..(2018). MENTAL MODELS ASSOCIATED TO VOICE USER INTERFACES FOR INFOTAINMENT SYSTEMS. DYNA, 93(3). 245. DOI: http://dx.doi.org/10.6036/8766

[21] Celaya-Padilla, J., Galván-Tejada, C., López-Monteagudo, F., Alonso-González, O., Moreno-Báez, A., Martínez-Torteya, A., Galván-Tejada, J., Arceo-Olague, J., Luna-García, H. and Gamboa-Rosales, H. (2018). Speed Bump Detection Using Accelerometric Features: A Genetic Algorithm Approach, Sensors 18(2): 443. Retrieved from http://dx.doi.org/10.3390/s18020443

[22] Nandal, A., Dhaka, A., Gamboa-Rosales, H., Marina, N., Galvan-Tejada, J., Galvan-Tejada, C., Moreno-Baez, A., Celaya-Padilla, J. and Luna-Garcia, H. (2018). Sensitivity and Variability Analysis for Image Denoising Using Maximum Likelihood Estimation of Exponential Distribution. Circuits, Systems, and Signal Processing, 37(9), pp.3903-3926.

[23] Gamboa Rosales, H. (2010). OptimaleBausteinauswahl in der KorpusbasiertenSprachsynthese. Dresden: TUDpress.

[24] Nematollahi, M., Vorakulpipat, C. and Gamboa, H. (2017). Digital Watermarking. 1st ed. Springer Singapore:.

[25] Gamboa Rosales, H., Jokisch, O. and Hoffmann, R. (2006). SPECTRAL DISTANCE COSTS FOR MULTILINGUAL UNIT SELECTION IN SPEECH SYNTHESIS. Proc. International Conference on Speech and Computer (SPECOM), pp.270-273.