ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

Personalizing Content Using Voice in a Digital Asset Ecosystem

Alexander I. Iliev, Peter Stanchev

Assistant Professor, Department of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

Professor, Department of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

ABSTRACT: Behind any cloud-based service there is a complex infrastructure that varies greatly depending on the industry and the types of services provided. Storing, searching and finding data through deep learning and artificial intelligence is the logical and necessary way forward. In the entertainment industry for example, digital media libraries offer massive amounts of media materials to the public where the big issues are finding, accessing and recommending specific content out of enormous set of choices. One way to approach the issue is through media-descriptive metadata, which comes as plain text (synopsis), sounds (narration), images (cover shots, etc.) or short videos (trailers). This however, is the conventional way that is building even more around the problem of finding specific content easily, hence not directly solving the main issue. This is so not only because specific applications must be developed, but also because they require physical human interaction with the system by using user dependent keystrokes in most cases. This in turn makes the access and extraction of digital content cumbersome and slow. Hence a better, more personalized automatic behind the scenes approach is needed.

KEYWORDS: Personalization, Emotion, Speech, Voice, Recognition, Digital Asset Ecosystem, Cloud Services

I. INTRODUCTION

The great fragmentation of digital collections, libraries and repositories puts on the agenda the question of providing users with opportunities for their joint consideration and study in order to fully utilize all semantic interconnections between them, overriding physical distance and the specifics of the digital storage of each source. One possible approach to solving this problem is linked to the creation of complex semantic-based and context-dependent models for improved use, research and delivery of large volumes of digital resources.

A most burning issue nowadays is the fundamental research for the context-sensitive (semantic-oriented) use of digitized content. Overcoming currently existing restrictions on the beneficial reuse of various digital repositories in different contexts and for different purposes is crucial for the development of the areas of Big data, Massive data mining, Data analytics, Data management and processing, Data visualization.The problem exacerbates even more when the world of open access knowledge increases its volume worldwide. Hence making use of all of these technologies can be assisted partially by innovations using digital signal-processing techniques and as a result will make their usage more precise and coherent.

II. THE ROLE OF DIGITAL SPEECH

To achieve the goals that lie ahead in any of the mentioned sub-fields, we first have to define some technological toolswith whichwe can obtainbetter results by processing, matching, selecting, recommending and personalizing of digital content. There could be many tiers to that extend depending on the feature domain.Collectingand processing of speech signals is one such tier. Speech is a convenient medium to analyze because it is nonintrusive and is easy to collect, process, transmit and store. Speech recognition is a vast area and includes many different layers such as: speech recognition – what has been said, speaker recognition – who said it, gender recognition – what was the gender of the speakeror emotion recognition – how it was said. For the general purposes the latter is of growing interest because emotions are much easier to convey, they are natural and do not carrysyntax. It is true however that conveying

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

emotions is very culture specific. That said, a system that correctly recognizes emotions can be trained to specific emotional patterns in different languages and in turn can be widely used. As a result, a number of services can benefit directly from possible nonintrusive automation based on speech. These services can vary all the way from: creating a suggestive search, recommendation or personalization of cloud based services in the digital media world, all the way to creating advancements in the higher education system.

For the purposes of this study we considered to centerthe speech analysis on its glottal layer alone. The shape of the glottal signal has been thoroughly investigated in the past and numerous models detailing its phases from a geometric prospective are known. For this study we chose the glottal model proposed by Fant [1–2]because it bringstheneeded analytical specifics of the shape of a typical glottal pulse. There are number of studies that focus on examining glottal features of speech under stress [3-7] and various glottal parameters have been investigated in emotion-related topics. Moore et al. [8,9] for example studied clinical depression using the spectral tilt and the glottal frequency response to derive inter-sentence and intra-sentence statistics.Ling et al. [10] used glottal information from the voice that was considered as an output of a Liljencrants-Fant model [11] working directly with glottal formant parameters and spectral tilt. In their case, glottal frequency characteristics were preserved in the speech spectrum and not obtained through inverse filtering.In our particular case we approach the problem by using pre-recorded American - English emotional speech database. More specifically, we perform closer examination of the glottal impulse and it's symmetry by applying inverse filtering methods on pre-recorded speech.

III. GLOTTAL ESTIMATION USING THE COVARIANCE METHOD

Having a glottal signal, obtained by direct recording from the larynx is easy and convenient by using a laryngograph. This however is only suitable in controlled laboratory environment. In real life,the only way to obtain aglottal signalfrom speech is through inverse filtering techniques. In order to fine-tune the algorithm exposed in this work, it is very beneficial to have both recorded and inverse filtered glottal signal. There are several methods that have been used to estimate the glottal signal. One such method is proposed by More and Clements (2003) [12]. The glottal closure (GCI) and glottal opening instances (GOI) are estimatedby using an iterative procedure applied on the raw speech. The results closely follow those provided with the aid of alaryngograph. Wong et al. (1979) [13]used a Linear Prediction covariance method to attain the residual signal. In it, the most negative peaks were depicted at the times of glottal closure. Using iteration the best estimate of the glottal waveform was obtained. For simplicity and to avoid ambiguity the glottal closure peaks were considered at midpoint of the glottal waveform. A shifting window with starting point locatedat each glottal closure was adopted when using the linear prediction (LP) in order to estimate the negative peaks.The length of the window was*2P*, which signified the amount of total shift around each glottal closure. The smoothest glottal waveform was obtained through applying a first order LP autocorrelation procedure on the glottal derivatives found from the iterative procedure. This is represented as follows:

$$\alpha_1 = \frac{-r(1)}{r(0)}, \qquad (1)$$

where, $\alpha_1 \approx 1$ when the glottal signalobtains thesmoothest shape. The numerator signifies the LP autocorrelation at *lag 1* and the denominator denotesthe autocorrelation at *lag 0*.

Figure 1 demonstrates the result obtained after applying the iterative glottal estimation mechanism by using the Linear Prediction autocorrelation algorithm on the main signal in order to obtain the glottal pulse. The stems represent the probable moments of glottal opening. On the plotting above we observe the residual signal and corresponding bellow is its extracted glottal counterpart.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

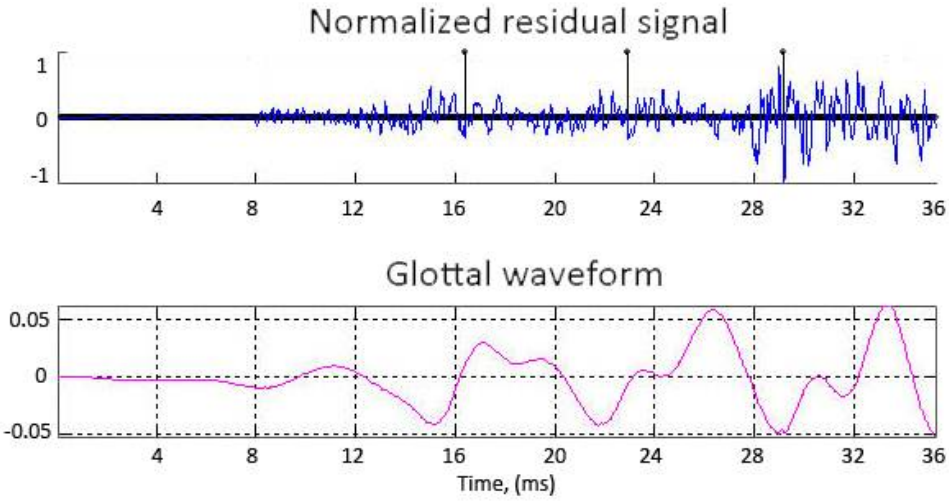

Fig.1: Glottal waveform obtained by using LP autocorrelation.

The processis illustrated in Figure 1. For the analysis it uses a window width of 36ms. As demonstrated, three iterations were completed in this example, each one corresponding to a glottal cycle.The iterationplacedan increased computational complexity on the procedure, providing however a more refined estimate of the glottal shape. It needs to be noted that the algorithm did not search for the exact moment of glottal closure and still produced a well-defined glottal shape. This fact alone showed that it was suitable for realistic real-life estimates of the glottal signal.

As established, by using iterations of the residual signal the best glottal signal can be obtained. However, some problems were encountered in certain instances when using the traditional covariance LPC method while trying to detect the exact GCI moments. In the cases when the prediction order *p* is high enough, the all-pole system can deliver good estimation, such that it will be equally good for the majority of speech samples. We must stress that this method used autoregressive LPC modeling, hencedelivering a good spectral estimate of the glottal flow signal. This however did not achieve a perfect deconvolution of the quasi-periodic glottal pulse train from the vocal tract function. There are a number ofissues that can stay on the way ofperfectreconstruction of the glottal signal when employing inverse filtering techniques. Despite of the errors that are introduced to the glottal signal, the glottal symmetry remainsin most cases impervious to fluctuation since it only considers theratio of the phases defined between the glottal openings and glottal closures, depicted by GCI and GOI. This makes the glottal symmetry domainextremely robust to various types of noise and thereforeproofs suitable for the emotion recognition problem defined in this work [14].

IV. AUTOCORRELATION LINEAR PREDICTION METHOD

As normal for any system, there are some limitations to be considered, which is why it is good to compare and evaluate further. The all-pole linear prediction method should be uniquely identified when previous samples from the output of the system are presented. It must be noted here that the need to adopt an all-pole assumption comes because of the fact thatwe do not have access to the incoming samples most of the time and consequently this model provides a system of equations that can be solved with high efficiency. A common all-pole system can be expressed as follows:

$$\tilde{y}(n) = \sum_{m=1}^{p} \alpha_m y(n-m) \quad , \qquad (2)$$

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

where, $\tilde{y}$ is the estimated output of the linear predictor, *p*signifies the prediction order, α_m represents the set of prediction coefficients and*y* is the true (actual) output of the system. The difference between the predicted or estimated signal and the actualsignal is expressed by the error defined below:

$$e(n) = y(n) - \tilde{y}(n), \quad (3)$$

In this study the prediction error is to be minimized by finding a better prediction estimation for the coefficients. To achieve that first we examined the performance of the methodusing*N* number of samples. Based on these observations we thenestablished the prediction order of the system *p*. Next, we calculated the predictor coefficients in a way that they had to minimize the energy of the error signal over *N* samples. This process led to solving a system of *p* equations with the same number of unknownsalso known as least-squares minimization.
The waveform signal used by the autocorrelation predictor, is considered to be bound within the interval [0, N-1]and can be expressed like this:

$$y(n) = y(n+k)\, w(n), \quad (4)$$

where, *n* is bound within the interval [0, N-1] and *w(n)*represents the finite length Hamming window. A window with *N* points was applied to the signal. The result of the liner predictor corresponds to the short-time autocorrelation function as:

$$\phi = \sum_{n=0}^{N-1-j+k} y(n)\, y(n-k+j), \quad (5)$$

where, $k \in [\,0, p\,]$and $j \in [\,1, p\,]$. The window boundaries determine that the signal is zero outside of the *N*-sample region. The autocorrelation expressionfrom equation (5) shapes a system of linear equations that can be represented asa matrix and can be solved using standard Gaussian elimination. The autocorrelation solution of these equations can be realized very efficiently because the autocorrelation coefficients in the matrix of equations have a very simple symmetric structure. This in turn allows for a recursive solution. Levinson and Durbin offer one such solution, where each predictor coefficient may be derived from the previous coefficient. The [*p* x *p*] matrix resolution of correlations in the autocorrelation LP method is a Toeplitz matrix that issymmetrical over the main diagonal such that all elements across are equal. This is why it is possibleto develop faster computational solution of this technique.

V. GLOTTAL ESTIMATION METHODOLOGY COMPARISON

The most commonly used linear prediction methodused for extraction of glottal signal is the covariance method. The autocorrelation LP method nonetheless carries certain advantagesin noisy conditions in Brooks et al. (2006) [15]. When comparing the two, there are three main issues to be addressed: 1) the number of multiplications in finding the solution of the matrix equation;2) the amount of storage needed for the matrix, and 3) stability of the performing method. Rabiner and Schafer (1978) [16] reviewed all three parameters in much detail:

1. The number of multiplications to calculate the correlation matrix in the covariance method is $* p$, and the solution of its equation requires $\frac{(p^3+9p^2+2p)}{6}$ multiplications, p divisions, and the same number of square roots. These numbers in the autocorrelation method are different. The later needs the same amount of multiplication for the correlation matrix that equal$N * p$, however the solution of the matrix equation uses much smaller number ofmultiplications orp^2.
2. The storage needed in the covariance method matches the number of analysis points*N* and for the correlation matrix that is $\frac{p^2}{2}$.These numbers for the autocorrelation method are: *N* for the data points and p for the autocorrelation matrix.They aresmaller than the one needed by the covariance method.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

3. Stability is directly related to the prediction order used and is very much guaranteed for the autocorrelation method when computed with sufficient accuracy. In addition, when using a pre-emphasis filter the stability of the predictor polynomials will normally remain stable. Stability of the prediction polynomials in the covariance method however cannot be guaranteed. Overall, both methods will lead to a similar solution if the number of samples in the analysis window is large enough. Considering the characteristics of both linear predictors discussed in this paper the autocorrelation remains in focus for this analysis.

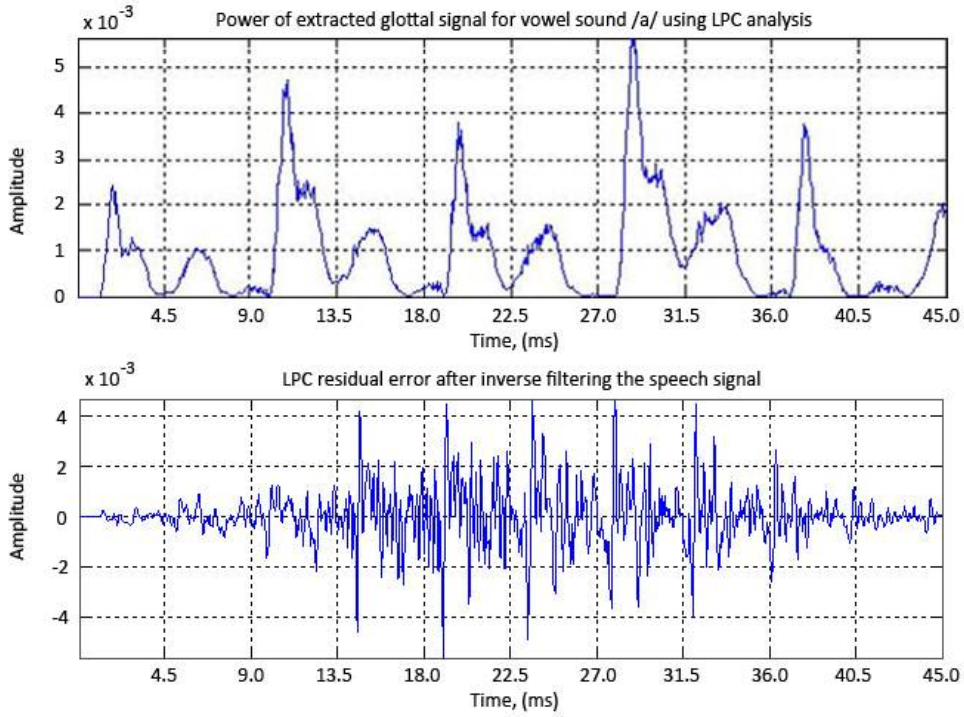

Fig. 2: Power glottal spectrum of vowel /a/ (above) and residual signal (below) after LPC inverse filteringof speech.

Figure 2 shows the power spectrum of extracted glottal signal through LPC analysis and an LP residual error of a given speech signal. The moments of glottal openings and closures in both projections are easy to see. The top portion of Figure 2 depicts a snipped of a power spectrum of the extracted glottal waveform for the spoken vowel sound /a/ from a short-time speech signal where the periodic nature of a glottal signal can be observed. The bottom portion illustrates the residual Linear Prediction error after the inverse filtering technique was applied. The peaks of the residual signal represent the moments of glottal closure. The two signals are not correlated.

VI. SPEECH CORPUS AND GLOTTAL SYMMETRY VALIDATION

The speech corpus used was created in an anechoic chamber with pre-recorded speech with American-English intonation containing four emotions conveyed in short and long emotional phrases. The signal was recorded with both a microphone and directly with the use of a laryngograph in orderto check the validity of the results when extracting the signal. In particular we were only interested in obtaining the glottal signal and calculating the open-close ratio of the glottisreferred to as Glottal Symmetry (GS) in Iliev et al. [17]. This is so because in a real world environment we only have access to raw speech, hence we need to adopt an inverse filtering method fort he extraction oft he GS.
The glottal signal is quasi-periodic in nature during voicing. It conveys important cluesof the identity of the speaker and their speaking style. In this study the role of the glottal signal was applied for the classification of spoken emotions.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

The glottal waveform of the first voiced section of any given speech utterance was estimated by usinginverse-filtering techniques [12,13,15].

In Iliev et al. (2006) [17] was established that the glottal information is rich in emotion evidences and it provides a very effective source for recognizing emotion in speech. Parameters such as glottal symmetry and its related tiermeasured at the beginning of a spoken utterance proved quite effective for classification. It was shown that glottal information was more effective than speech alone and any combinations of glottal and speech features slightly degraded the overall performance.

In [18] a sequential covariance 12th order LP error is used, and comparedalongside the laryngograph-recorded and inverse filtered glottal signals. In this work was confirmed that the recorded and inverse filtered glottal signals have similar features. Moreover, an assessment between the recorded speech analysis and the synthesized one was included andit was established that there are important parallels between the two signals extracted via inverse filtering. More importantly, the Rosenberg-type excitation pulse (1971) [19] used for the synthesis achievednear perfect reconstruction.

During testing we encountered problems in some voiced regions where both algorithms failed to represent the glottal formaccurately. This was likely due to the all-pole assumption of the vocal tract. Simply put, for some speakers this assumption does not performedas well for voiced consonants as when compared to vowels sounds. More specifically, in the production of nasal consonants such as (m, n) the oral cavity is closed, which introduces zeros in the vocal tract response. This is the reason why the vocal tract prototype does not reflect the realistic physical model because some glottal closures are notprotuberant in the residual signal.In some extreme cases, when working with nasal consonants, there may be a problem in determining the exact moments of closure.This is due to smearingwhen using the conventional LPC analysis.

In the corpus used here, the performance of seven different classifiers was compared. One such classifier was the Optimum-Path Forest (OPF) classifier [20]. It was compared against six other methods namely: ANN-MLP, BC, C4.5, GMM, k-NN, SVM.

Figure3 shows a graphical summary of the comparison results, where can be seen that the best classification performance was achieved by SVM and OPF and the lowest performance was provided by GMM. As far as computational times, k-NN showed the fastest performance. It is important to note that from the top two classifiers, OPF was considerably faster than SVM, thus making the former quite attractive for theproblem at hand. These results were based on experiments with various feature vectors shown on the x-axis, which are: GS – glottal symmetry, 8(gl)/10(gl) – 8/10 glottal features respectively, 8(sp)/10(sp) – 8/10 classical speech parametersas defined in [21, 22].

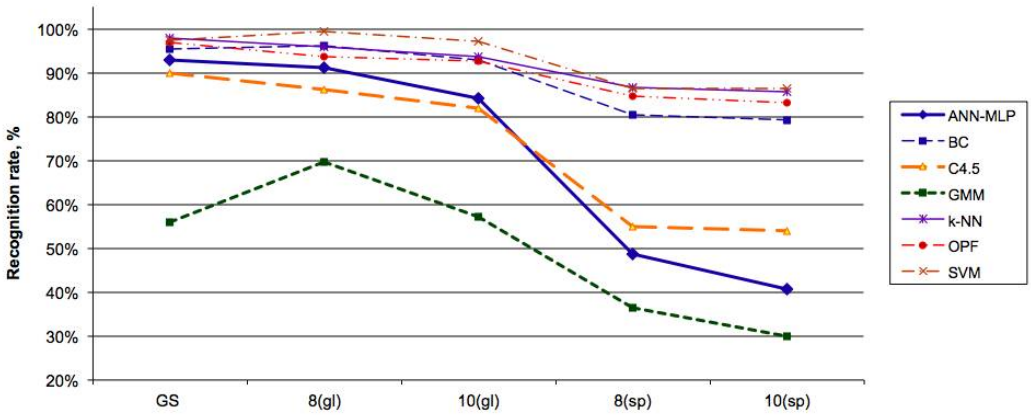

Fig.3: Mean recognition rates for 5feature vectors using Angry, Happy, Sadand Neutralemotions in 10 rounds.

VII. DISCUSSION OF RESULTS

The average performance of correctly recognized emotion patterns from the corpus used in this study was 81.05%. Detecting the correct emotional state for all speakers combinedwas done while using all features, emotions and classifiers collectively. One probable reason for the high recognition rate achieved was because the speech was

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

recorded in an anechoic chamber and therefore the obtained speech quality was higher, hence no noise was present in the recording. The prerecorded glottal signal used for reference, also played a big role for comparison of the results.

When comparing the results we observed thatthe difference between Sad and Neutral are generally miniscule, as they both naturally exhibit a 'low-key' behavior. The difference between the more energetic expression of emotion in the 'Happy' and 'Angry' sets is visible as compared with the first two. One can also observe the difference in GS open-close ration between 'Happy' and 'Angry'. The latter exhibits more spices in the analyzed domain. We also see that the ratios of the more saddle emotions of interest 'Neutral' and 'Sad' is higher. This is due to a smoother opening phase in relation to the one for 'Angry' and 'Happy', which is forced and faster as a result.

VIII. CONCLUSIONS

In this work was shown that emotional recognition of speech signals and in particular for the feature parameters chosen here could be of great benefit to searching, finding and recommending of personalized content of any kind. This is because the detection of emotional patterns is distinctive enough so that it can be connected to personal preferences based on content use and consumption founded on emotion. Moreover, it can be successfully applied to many other solutions that include, but are not limited to media data selection in the entertainment industry, media metadata description and added to any media provider cloud service to ease the recommendation process, it can be applied for security purposes in major public places such as airports, train stations, city sights, etc., it can be applied to any hospital monitoring scenarios where with the assistance of technology services and care for patients may be improved by, it can be holistically applied to any of the problems outlined at the beginning of this work, namely to Big data, Massive data mining, Data analytics, Data management and processing, Data visualization. Finally, the findings displayed in this work show that emotional recognition of speech signals can be successfully applied to any cloud hosting solution that require some level of personalization. Further in-depth research may be conducted with an emphasis on gender separation in making the personalization more accurate and robust to variations.These findings can be viewed as part of a larger digital asset ecosystem.

REFERENCES

[1] G. Fant, "The voice source—acoustic modeling," Speech Transmission Laboratory, Quarterly Progress and Status Reports, no. 4, pp. 28–48, 1982
[2] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," Speech Transmission Laboratory, Quarterly Progress and Status Report, no. 4, pp. 1–13, 1985
[3] K. E. Cummings and M. A. Clements, "Analysis of glottalwaveforms across stress styles," in Proceedings of the IEEEInternational Conference in Acoustics, Speech, and SignalProcessing, pp. 369–372, April 1990
[4] K. Cummings and M. Clements, "Improvements to and applications of analysis of stressed speech using glottal waveforms," in Proceedings of the IEEE International Conference in Acoustics, Speech and Signal Processing, pp. 25–28, 1992
[5] K. E. Cummings and M. A. Clements, "Application of the analysis of glottal excitation of stressed speech to speaking style modification," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 207–210, April 1993
[6] K. E. Cummings and M. A. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech," Journal of the Acoustical Society of America, vol. 98, no. 1, pp. 88–98, 1995
[7] A. M. Laukkanen, E. Vilkman, P. Alku, and H. Oksanen, "Physical variations related to stress and emotional state: a preliminary study," Journal of Phonetics, vol. 24, no. 3, pp. 313–335, 1996
[8] E.Moore,M. Clements, J. Peifer, and L.Weisser, "Investigating the role of glottal features in classifying clinical depression," in Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2849–2852, September 2003
[9] E.Moore,M. A. Clements, J.W. Peifer, and L.Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," IEEE Transactions on Biomedical Engineering, vol. 55, no. 1, pp. 96–107, 2008
[10] Z. H. Ling, Y. Hu, and R. H. Wang, "A novel source analysis method by matching spectral characters of LF model with STRAIGHT spectrum", in Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction (ACII '05), vol. 3784, pp. 441–448, 2005
[11] G. Fant, "Glottal flow: models and interaction," Journal of Phonetics, vol. 14, pp. 393–399, 1986
[12] Moore E., Clements M., Peifer J., and Weisser L., "Investigating the Role of Glottal Features in Classifying Clinical Depression", 25th Annual International Conference of the IEEE EMBAS, pp. 2849-2852, 2003
[13] Wong D., Markel J., and Gray A, "Least Squares Glottal Inverse Filtering from the Acoustical Speech Waveform", IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-27, No. 4, pp. 350-355, 1979
[14] Iliev, A.I., Scordilis, M.S., "Spoken Emotion Recognition Using Glottal Symmetry", EURASIP Journal on Advances in Signal Processing, Volume 2011, Article ID 624575, 2011

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

[15] Brooks M., Naylor P., and Gudnason J., "A Quantitative Assessment of Group Delay Methods for Identifying Glottal Closures in Voiced Speech", IEEE Transactions On Audio, Speech and Language Processing, Vol. 14, No. 2, pp. 456-466, 2006
[16] Rabiner L. and Schafer R., "Digital Processing of Speech Signals", Prentice Hall, 1978
[17] Iliev, A.I., Scordilis, M.S., "Emotion Recognition in Speech using Inter-Sentence Glottal Statistics", Proceedings of the 15th International Conference on systems, Signals and Image Processing, IEEE-IWSSIP 2008, Bratislava, Slovakia, pp. 465-468, June 25-28, 2008
[18] Iliev, A. I., "Emotion Recognition From Speech", Monograph, Lambert Academic Publishing, 2012
[19] Rosenberg A., "Effect of Glottal Pulse Shape on the Quality of Natural Vowels", Journal of the Acoustical Society of America, Vol. 49, pp. 583-590, 1971
[20] Iliev, A.I., Scordilis, M.S., Papa J.P., Falcão A.X., "Spoken emotion recognition through optimum-path forest classification using glottal features", Journal on Computer Speech and Language, ELSEVIER, Vol. 24, Issue 3, pp. 445-460, 2010
[21] Iliev, A.I., Zhang, Y., Scordilis, M.S., "Spoken Emotion Classification Using ToBI Features and GMM", Proceedings of the 14th International Workshop on Signals and Image Processing 2007 and the 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. IEEE-IWSSIP 2007, Maribor, Slovenia, pp. 495-498, June 27-30, 2007
[22] Iliev, A.I., "Emotion Recognition in Speech using Inter-Sentence Time-Domain Statistics", IJIRSET International Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, Issue 3, pp. 3245-3254, March 2016