# *Cries and Whispers*
# Classification of Vocal Effort in Expressive Speech

*Nicolas Obin* †

IRCAM-CNRS UMR 9912-STMS
Paris, France
nobin@ircam.fr

## Abstract

The expansion of the video games industry raises innovative and challenging issues for speech technologies, e.g. the development of automatic content-based speech processing and speech recognition systems in the context of video games post-production and voice casting. This paper presents a large-scale study on the classification of vocal effort in expressive speech for video games. Changes in vocal effort conduct to substantial modifications in the configuration of voice production mechanisms. In particular, registers of vocal effort affect especially voice quality which reflects qualitative modifications of the source excitation characteristics. This study introduces robust source characteristics to measure various types of voice quality (e.g., breathy, creaky, tense) for the classification of vocal effort into whispered, normal, and shouted speech. The system is evaluated in the real scenario of video games production with the complete speech recordings of a massive role-playing video game. The proposed features significantly improve the classification from 81.1% to 87% over conventional MFCCs. These advancements confirm the role of the source and voice quality for the description of changes in vocal effort.
**Index Terms** : speech recognition, vocal effort, voice quality, glottal source, GMM-UBM/SVM.

## 1. Introduction

Vocal effort corresponds to the adjustment of vocal intensity of a speaker depending on the communication distance to the listener. Vocal effort may also vary with respect to various causes (e.g., degree of intimacy that relates speaker and listener, emotional state of the speaker). A change in vocal effort causes a change in vocal intensity which conduct to substantial modifications in the configuration of voice production mechanisms. Studies on vocal effort usually assumes 5 configurations of voice production : whispered, soft, normal, loud, and shouted speech [1].

A large number of studies has been dedicated to the description of the mechanisms involved in the production of whisper [1, 2, 3], and shouted [4] speech, and the differences in configuration between soft, normal, and loud speech [5]. *True* whisper (low-effort whisper) refers to the excitation of the vocal tract with half-closed vocal folds. *Stage* whisper (high-effort whisper) is the simulation of whisper speech by professional actors so the speech is sufficiently loud so that the speech is audible by the audience. Stage whisper differs from true whisper with extremely breathy and partially voiced speech [2]. Shouted speech corresponds to a raise in F0 due to the raise of subglottal pressure used to increase vocal intensity [6], and non-linearities due to the non-linear interaction of air flow and air vortices near the vocal folds producing additional excitation signals. A raise in vocal effort also affects the characteristics of the vocal tract, from the muscle tension in the vocal tract to the modification of vocal tract resonances. Finally, these modifications can be retrieved from signal analysis and the observation of F0 and spectral characteristics of speech [7] (e.g., spectral slope as an approximation to the glottal source structure).

The identification of vocal effort drastically affects the performance of speech recognition systems, from speech segmentation to speaker identification [8, 9, 10]. Vocal effort identification systems has been proposed to adapt speech recognition systems to specific vocal efforts [11, 12, 13, 14, 8, 10]. In the context of video games production, the classification of vocal effort configuration is critical for audio post-production (e.g., application of specific settings for speech level and multi-band compression depending on the vocal effort in order to simulate the proximity of speaker and listener), and may also helps for automatic voice casting. Additionally, the context of video games raises challenging issues compared to the requirements of conventional vocal effort identification systems, from the large database ($\simeq$ 20.000 to 40.000 audio files and $\simeq$ 500 roles), the large range of speech variability of professional actors, including unnatural speech (e.g., cartoons, robots, extraterrestrials) ; the large difference in audio recordings duration from a single filler (e.g., inspirations, screams, $\leq 0.5s.$) to complete utterances ($\simeq 10s.$).

This study presents a GMM-UBM/SVM system [15] and investigates robust source excitation characteristics to measure various types of voice quality [16] for the classification of vocal effort into whispered/soft, normal, and loud/shouted speech. In particular, this study will exploit recent advances in the robust separation of glottal source and vocal-tract filter [17, 18] for speech recognition systems. The proposed source characteristics are : soft voiced/unvoiced description (VUV) in Mel-frequency bands as a measure of *breathiness* ; the glottis relaxation coefficient (RD) as a measure of *tension* in the voice ; and the irregularity of the glottal pulse ($\Delta$GCI) as a measure of *creakiness/breathiness*. The proposed source characteristics are compared with conventional features - MFCCs and Teager Energy Operator (TEO) [11, 12] - within a large-scale evaluation in the real scenario of video games production conducted on the complete speech recordings of a massive role-playing video game.

†This study was supported by the European FEDER project VOICE4GAMES.

1. Lombard and stressed speech may also be additional configurations.

## 2. Speech Database

Speech recordings of a video game consist in the interpretation of script lines by professional actors who are directed by an artistic head whose role is to control the expressive content of speech depending on the place of the script within the overall scenario. Script lines may vary from a single sigh to a complete sentence. In role-playing games, the recording covers ten-thousands of speech files that are split into hundreds of roles. The video game used for the study includes around 20.000 French speech recordings and 500 roles (from a single to hundreds of recordings). The duration of speech recordings varies from 0.1 seconds to 20 seconds with a mean duration of 2.5 seconds. Recordings were made in mono 48 kHz/16 bits uncompressed format.

Speech recordings are produced in a studio by professional actors with a varying distance and orientation to the microphone so as to compensate the variations in acoustics due to changes in intended vocal effort (close while whispering, distant while shouting). Additionally, a sound engineer ensures that the speech level is constant through speech recordings so as to provide a homogeneous speech level through the video game. Finally, the situation of professional studio recording of speech for video games exhibits no significant differences of speech level for changes in vocal effort by professional actors (contrary to laboratory recordings [12]). Hence, only information provided by changes in the source excitation or the vocal tract resonances can be used for the identification of changes in vocal effort.

The identification of significant changes in vocal effort is critical in the production of video games for the application of specific settings that simulate the perception of proximity to the speaker by the player. For this purpose, the sound engineer is usually in charge for the manual classification of expressive speech recordings into three classes which cover whispered/soft, normal, and loud/shouted speech. In the present study, whispered/soft speech covers sighs, true whisper, stage whisper, stressed whisper (tense whisper typically produced in a life-survival situation in which the conversation intimacy is absolutely required), soft speech, and any situation of intimacy in the speech communication. Loud/shouted speech includes orders, public announcements, exclamations, interjections, stressed-speech, and screams.

## 3. Source Characteristics

The objective of this study is to investigate robust speech characteristics which may provide complementary information to the conventional MFCCs. Robust characteristics denote characteristics that are not subject to errors which may be critical for the classification. For instance, F0 and VOICED/UNVOICED FREQUENCY (VUF) were not retained for the comparison due to their sensibility in the context of expressive speech - especially, for the analysis of whispered and shouted speech.

Instead, robust source characteristics are introduced that may relate to various types of voice quality (breathiness, creakiness, tense, stress). For instance, a measure for the description of noisiness in the speech is introduced as a robust reformulation of the VUF into a soft VOICED/UNVOICED DECISION (VUV) within Mel-frequency bands. Additionally, this study will exploit recent advances in the robust separation of glottal source and vocal-tract filter [17, 18] which have been proved to be robust for the analysis of expressive speech. In particular, the glottal-source and vocal-tract separation provides an explicit description of the glottal source characteristics contrary to methods which does not process this separation (e.g., spectral slope). The glottal source characteristics considered are : the glottis relaxation coefficient (RD) as a measure of *tension* in the voice ; and the irregularity of the glottal pulse ($\Delta$GCI) as a measure of *creakiness*. Finally, TEAGER ENERGY OPERATOR (TEO) based characteristics are also used for comparison which were proved to be extremely relevant for the identification of loud and stressed speech [12].

### 3.1. MFCC

13 Mel-Frequency Cepstral Coefficients (MFCC) are extracted after the non-linear compression of amplitude spectrum into 25 Mel-frequency bands. Short-term features were extracted with a 25 ms. hanning moving window of 5ms.

### 3.2. TEO

The Teager Energy Operator has been introduced as a measure to reflect the non-linear airflow structure of speech production that can be observed in emotional and stressed speech [11, 12]. The non-linearities introduced in the source excitation causes the apparition of additional partials and modulations in the spectrum which do not correspond to the usual linear source/filter model. The TEO profile is a constant in the presence of a single and stationary sinusoid. This property can be extended so as to measure separately the degree of non-linearities within various frequency bands. Further investigations conduct to the introduction of the normalized TEO autocorrelation envelope area which intends to reflect the distribution, the degree of stationarity, and the interactions of partials that are present within a frequency region. The normalized TEO autocorrelation envelope area has been proved to be extremely consistent for the identification of loud and stressed speech [12]. In the present study, the normalized TEO autocorrelation envelope area was extracted after band-pass filtering of the speech signal into 25 Mel-frequency bands.

### 3.3. V/UV

The frequency distribution of harmonic content is assumed to change significantly with raise in vocal effort and reflect the degree of vibration of the vocal folds, the vocal intensity, and the raise of F0. In particular, a raise in vocal effort may coincide with a raise in the harmonic distribution through high-frequency bands. In this study, a soft Voiced/Unvoiced decision (VUV) [19] is introduced as a measure of the harmonic distribution into 25 Mel-frequency bands. For each frequency band, the VUV is measured as :

$$VUV^{(i)} = \frac{\sum_{k=1}^{K^{(i)}} |A_H(k)|^2}{\sum_{n=1}^{N^{(i)}} |A(n)|^2} \quad (1)$$

where : $i$ denotes the $i$-th Mel frequency band, $A_H(k)$ the amplitude of the $k$-th harmonic [20], and $A(n)$ the amplitude of the $n$-th frequency bin in the considered frequency band. Hence, VUV is equal to zero when no harmonic content is present in the frequency band, and to one when only harmonic content is present in the frequency band.

Figure 1 presents the distribution of the VUV for whispered/soft, normal, and loud/shouted speech. This clearly shows

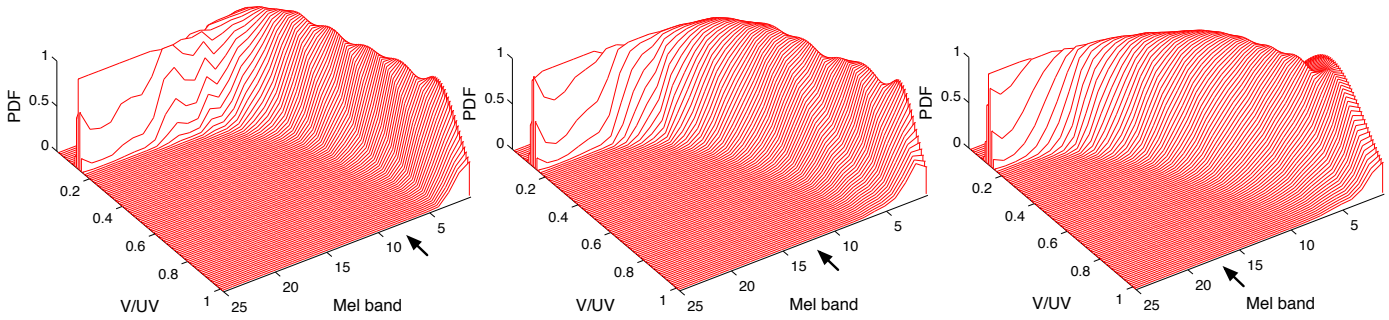


FIGURE 1 – Distribution of the Mel-frequency V/UV for whispered, normal, and shouted speech - from left to right. Black arrows indicate the Mel-frequency band upon which no voiced frequency content is observed (e.g., $f_W = 9$, $f_N = 14$, and $f_S = 18$ for $V/UV \leq 0.2$)

that the raise of vocal effort conduct to a significant raise of the harmonic content through high frequency bands. For instance, the frequency region upon which no significant harmonic content is present raises from the 9-th frequency band for whispered/Soft speech to the 18-th frequency band for loud/shouted speech.

### 3.4. Voice Quality

Glottal source characteristics are extracted from the robust separation of glottal source and vocal-tract filter [17, 18]. In particular, the glottal shape (LF-RD [21]) and Glottal Closure Instants (GCI) are determined separately. In this representation, the Liljencrants-Fant (LF) glottal model is described by a single parameter referred as the glottal relaxation coefficient (RD) [18]. Then, the estimated Rds are used to determine the position of the GCIs [17]. Then, the regularity of the GCIs is determined as :

$$\Delta GCI(n) = |\Delta^2 \log(GCI(n) - GCI(n-1))| \quad (2)$$

where : $\Delta^2(.)$ denotes the second derivative operator, and $GCI(n)$ the $n$-th GCI time position.

These characteristics are referred in this study as voice quality characteristics (VQ) since there are explicit determination of the glottal source characteristics which relate to qualitative voice qualities : RD provides a description of the degree of *tension* in the voice, and the regularity of the GCIs provides a description of the degree of *creakiness* and/or *breathiness* in the voice.

Figure 2 presents the distribution of the VQ characteristics (RD, $\Delta$ GCI) for whispered/soft, normal, and loud/shouted speech. This clearly shows that the raise of vocal effort conduct to a significant differences in the configuration of the glottal source characteristics. In particular, a raise in vocal effort corresponds to a raise in tension of the vocal tract. Moreover, whispered/soft speech exhibits a large dispersion in the glottal pulse regularity which clearly reflects the presence of whispery/breathy/creaky speech. Also, normal speech exhibits a larger dispersion in the glottal pulse regularity than loud/shouted speech which may be due to the presence of intrinsically breathy/creaky speakers in normal speech.

## 4. Evaluation

The relevance of the proposed speech characteristics for the classification of vocal effort into whispered/soft, normal, and loud/shouted speech has been conducted within a 5-fold cross-validation. Additional constraints on the design of the cross-validation have been adopted : well-balanced distribution of the vocal effort classes within the train and test sets ; no role overlapping across train and test sets in order to prevent the system to turn into a speaker identification system.

The classification system is based on the GMM-UBM/SVM system [15] which is a standard for speech recognition [22, 23]. A UNIVERSAL BACKGROUND MODEL (GMM-UBM) is used to model the acoustic variability in the speech database with a Gaussian Mixture Model (GMM). Then, each utterance is represented as a SUPERVECTOR by MAP adaptation of the GMM-UBM mean vectors [22]. Then, SUPERVECTORS are used to determine the parameters of a SUPPORT VECTOR MACHINE classifier which maximize the margin of a high-dimensional separation hyperplane for the classification [23].

During the training, each feature set is considered as a separate stream for the determination of the GMM-UBM and the SVM parameters. 64 GMMs with diagonal covariance matrices have been used to determine the parameters of the GMM-UBM, and a SVM with a GMM-supervector Radial Basis Function (RBF) kernel has been determined with various values of the radial bandwidth (from 0.1 to 5). During the classification, the decision is made by fusioning the affinity obtained for each stream using average decision fusion. The performance of the system has been measured with the F-measure metric. Finally, the performance of the system corresponds to the optimal performance obtained for each feature set.

Table 1 summarizes the system performance obtained for conventional MFCC, combination of introduced feature sets with MFCC, and the optimal combination of feature sets for the classification of vocal effort.

| **MASS EFFECT** | WHISPERED | NORMAL | SHOUTED | TOTAL |
|---|---|---|---|---|
| MFCC | 69.4 | 82.5 | 91.0 | 81.1 |
| MFCC + TEO | 72.5 | 83.6 | 91.6 | 82.5 |
| MFCC + VUV | 75.0 | 84.3 | 91.2 | 83.5 |
| MFCC + VQ | 76.9 | 84.7 | 92.0 | 84.7 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| MFCC + TEO + VUV + VQ | 79.1 | 88.4 | 93.5 | 87.0 |

TABLE 1 – F-measure obtained for the conventional MFCC, combination of proposed features with MFCC, and optimal configuration for the MASS EFFECT role-playing video game.

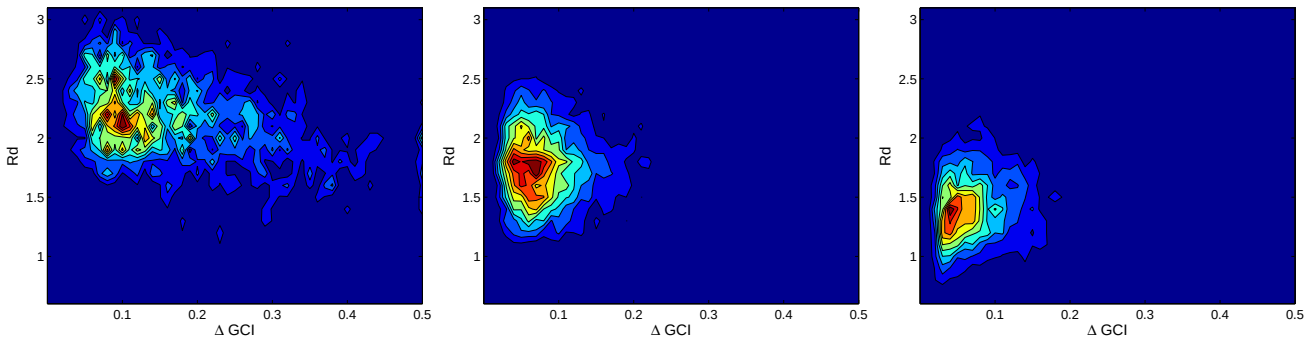


FIGURE 2 – Distribution of the voice quality characteristics ($\Delta$ GCI, $R_d$) for whispered, normal, and shouted speech - from left to right.

The evaluation exhibits that the proposed source characteristics carry complementary information for the classification of vocal effort. The VQ characteristics are proved to be extremely relevant for the classification of vocal effort (from 81.1% to 84.7%), especially for the classification of whispered/soft speech (from 69.4% to 76.9%). Also, the proposed VUV and VQ characteristics significantly outperform the conventional MFCC and TEO characteristics. Finally, the optimal performance is obtained with the complete combination of investigated features (87.0%), which indicates that each feature set carry complementary information for the classification of vocal effort.

## 5. Conclusion

In this paper, a large-scale study on the classification of vocal effort in expressive speech was presented. This study introduced robust source characteristics to measure various types of voice quality (e.g., breathy, creaky, tense) for the classification of vocal effort into whispered/soft, normal, and loud/shouted speech. The system is evaluated in the real scenario of video games production with the complete speech recordings of a massive role-playing video game. The proposed features significantly improve the classification from 81.1% to 87% over conventional MFCCs. These advancements confirm the role of the source excitation and voice quality for the description of changes in vocal effort. In further studies, the robustness of vocal effort classification will be assessed with cross video games and cross languages evaluations.

## 6. References


[1] P. Monoson and W. R. Zemlin, "Quantitative Study of a Whisper," *Folia Phoniat*, vol. 36, no. 2, p. 53–65, 1984.

[2] N. P. Solomon, G. N. McCall, M. W. Trosset, and W. C. Gray, "Laryngeal Configuration and Constriction during Two Types of Whispering. ," *Journal of Speech and Hearing Research*, vol. 32, pp. 161–174, 1989.

[3] J. Sundberg, R. Scherer, M. Hess, and F. Müller, "Whispering - A Single-Subject Study of Glottal Configuration and Aerodynamics," *Journal of Voice*, vol. 24, no. 5, pp. 574–584, 2010.

[4] D. Rostolland, "Acoustic Features of Shouted Voice," *Acustica*, vol. 50, pp. 118–125, 1982.

[5] E. B. Holmberg, R. E. Hillman, and J. S. Perkell, "Glottal Airflow and Transglottal Air Pressure Measurements for Male and Female Speakers in Soft, Normal, and Loud Voice," *Journal of the Acoustical Society of America*, vol. 84, no. 2, pp. 511–529, 1988.

[6] P. Gramming, J. Sundberg, S. Ternström, R. Leanderson, and W. H. Perkins, "Relationship between Changes in Voice Pitch and Loudness," *Journal of Voice*, vol. 2, no. 2, p. 118–126, 1988.

[7] C. Harwardt, "Comparing the Impact of Raised Vocal Effort on Various Spectral Parameters," in *Interspeech*, Florence, Italy, 2002, pp. 2941–2944.

[8] C. Zhang and J. H. L. Hansen, "Analysis and Classification of Speech Mode : Whispered through Shouted," in *Interspeech*, Antwerp, Belgium, 2007, pp. 2289–2292.

[9] ——, "Effective Segmentation based on Vocal Effort Change Point Detection," in *Speech Analysis and Processing for Knowledge Discovery*, Aalborg, Denmark, 2008.

[10] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of Vocal Effort Variability on Automatic Speech Recognition," *Speech Communication*, vol. 54, no. 6, p. 732–742, 2012.

[11] H. M. Teager and S. M. Teager, "Evidence from Nonlinear Production Mechanisms in the Vocal Tract," *Speech Production and Speech Modeling*, vol. 55, pp. 241–261, 1989.

[12] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear Feature Based Classification of Speech Under Stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.

[13] S. J. Wenndt, E. J. Cupples, and R. M. Floyd, "A Study on the Classification of Whispered and Normally Phonated Speech," in *International Conference on Spoken Language Processing*, Denver, Colorado, 2002, pp. 649–652.

[14] T. Ito, K. Takeda, and F. Itakura, "Analysis and Recognition of Whispered Speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, 2010.

[15] C. Charbuillet, D. Tardieu, and G. Peeters, "GMM-Supervector for Content based Music Similarity," in *International Conference on Digital Audio Effects*, Paris, France, 2011, pp. 425–428.

[16] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge : Cambridge University Press, 1980.

[17] G. Degottex, A. Roebel, and X. Rodet, "Glottal Closure Instant Detection from a Glottal Shape Estimate," in *13th International Conference on Speech and Computer*, St-Petersburg, Russia, 2009, pp. 226–231.

[18] ——, "Phase Minimization for Glottal Model Estimation," *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 19, no. 5, pp. 1080–1090, July 2011.

[19] D. W. Griffin and J. S. Lim, "A New Model-Based Analysis/Synthesis System," in *International Conference on Acoustics, Speech, and Signal Processing*, Tampa, Florida, 1985, pp. 513–516.

[20] M. Zivanovic, A. Röbel, and X. Rodet, "Adaptive Threshold Determination for Spectral Peak Classification," *Computer Music Journal*, vol. 32, no. 2, pp. 57–67, 2008.

[21] G. Fant, "The LF-Model Revisited. Transformations and Frequency Domain Analysis," K.T.H. Quarterly Progress Report and Status Progress. Departement for Speech, Music and Hearing, Tech. Rep. 2-3, 1995.

[22] R. B. D. Douglas A. Reynolds, Thomas F. Quatieri, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[23] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.