*ISSN 1392 - 1215 ELEKTRONIKA IR ELEKTROTECHNIKA. 2003. Nr. 6(48)*

*T121 SIGNALŲ TECHNOLOGIJA*

# Aspects of Melisma Synthesis

**R. Leonavičius, D. Navakauskas**
*Radioelectronics Department, Vilnius Gediminas Technical University*
*Naugarduko 41-422, LT-2006 Vilnius, Lithuania, tel. +370 5 2744756, e-mail dalius@el.vtu.lt*

## Introduction

A field of speech signal synthesis having various application areas is now in its mature. Still main application area is in the telecommunication when speech synthesis is used mainly as a mean to reduce throughput of communication channels and improve intelligibility. However times past, when the only objective of speech synthesis was intelligibility. Worth to notice is the need to have various systems that could communicate with people effortless, e.g., using synthetic speech. That requires synthesizing not only intelligible but also perceptually indistinguishable artificially generated speech, possessing such features as intonation, accent, speaker introduced personal variations and even emotions.

Our applications target is the restoration of audio signals recorded in old records [1, 2]. We look to a restoration problem from a quite different perspective, considering a speech synthesis as a mean of restoration of long audio fragments. Thus similarly requirements as previously mentioned of the quality of restored segments could be stated. The problem becomes especially difficult, when someone aims to restore/synthesize songs. Perception of song is not only dependent on the auditory, but also on performer skills and maturity, acoustic characteristics of room (compare office environment with orchestra hall), etc. Being so complicated, here we address song synthesis problem only partially – in this paper we deal with the synthesis of melisma [3], a common feature of any song, showing initial restoration results based on the several case studies.

Modern synthesis of speech signals [4] essentially is based on utilization of huge volume vocabularies (ScanSoft RealSpeak™ system [5, 6]) or sets of phonemes (The Bell Labs TTS system [7], The Festival Speech Synthesis System [8]), or a mixture of both of them (Hybrid ORATOR® II Speech Synthesizer by Telcordia [9], ProVerbe Speech Unit™ by Elan Speech [10]). However, mentioned methods do not conceptually fit in a framework of the synthesis of melisma due to unlimited number of variations that it possesses, whilst in essence these variations are the key feature of melisma and it does make perception of it beautiful.

The paper is organized as follows. Introduction of melisma from the engineering point of view, i.e., enlightening and presenting main characteristics such as intensity, pitch and spectrogram is given at the beginning. Presentation of several preliminary ways of synthesis of melisma is given next. First initial experiment is done in a way that enables us to synthesize melisma perfectly because of natural characteristics of melisma are employed. Second experiment set-up explores synthesis of melisma employing artificially generated (according to deterministic expression) pitch signal. The third experiment and its set-up expresses our suggestion and main contribution of the paper – use of the segment of original glottal impulse in order to synthesize melisma. The paper ends with main ideas summarized in conclusions.

## Melisma and its characterization

Melisma (from Greek – *melody*) is an expressive vocal phrase or passage consisting of several nodes sung to one syllable [11]. From the point of view of signal theory, melisma is a non-stationary signal, i.e., characterized by rapidly changing amplitude (intensity) and periodicy (pitch). Melisma are classified into four main groups [3, p. 274] called *trill*, *gruppett*, *fortis* and *mordent*. For the following study we select *gruppett* as an object of interest. Particular realization of *gruppett* as a waveform is shown in the Fig. 1a, while calculated characteristics such as pitch shown by thick solid line and intensity indicated by thin line are shown in Fig. 1b and spectrogram is shown in Fig. 1c.

The data presented in Fig. 1 confirm the non-linear nature of characteristics while *gruppett* is performed. It is seen, that pitch changes quite severe, i.e., from 370 Hz to 270 kHz at the same time intensity variability is also big. Most modern systems do not let to incorporate such variability of pitch and intensity at least in the word ranges. Spectrogram could serve as a composite mean to show the complicated and rapid changes in a signal when *gruppett* or actually all kinds of melisma are performed. Thus stipulates the need to select adequate method that could capture rapid variability of the signal. In the following we are going to employ method that will utilize waveform information through integrated characteristics and their approximations.

## Synthesis of melisma using its natural parameters

Based on the classical approach, here we synthesize melisma while employing signal of glottal impulses and vocal tract model [12]. We use Linear Prediction Coding (LPC) method, during which we estimate coefficients of $k$ order linear autoregression process (AR) model

$$\hat{x}(n) = -a_2 x(n-1) - a_3 x(n-2) - \ldots - a_{k+1} x(n-k). \quad (1)$$

Here $x(n-k)$ is a signal instance at discrete time $n$, estimate of it is denoted as $\hat{x}(n)$, while model coefficients are denoted by $a_k$.

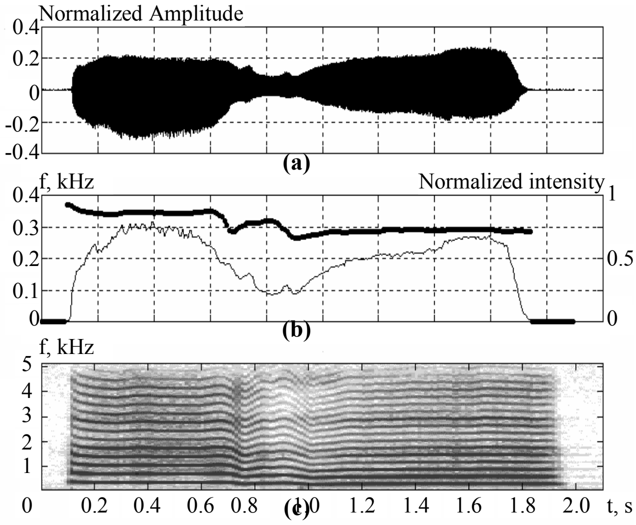


**Fig. 1.** Characteristics of a *gruppett*: (a) waveform, (b) pitch (thick line) and intensity (thin line), (c) spectrogram calculated using 1/2 overlapping 1024 points Hanning window FFT

Length of a synthesis frame (or assumed time for stationarity of the signal) is kept as in most synthesis systems 30 ms, while frame is moved by 10 ms duration steps.

The overall procedure of melisma synthesis is as follows. Initially, signal passes amplitude detector (voiced/unvoiced decision) that eliminates unvoiced parts at the beginning and end of particular melisma. Then, by the means of LPC, coefficients of the AR model are estimated. Using estimates of coefficients, inverse model is constructed, enabling to determine signal of original glottal pulses. Having decomposed particular melisma into its counterparts, synthesis using natural (original) signal of glottal pulses and instantaneous vocal model is straightforward. We mention here only the need to use smoothing window (in our case Hamming window) in order to reduce rectangular window effects, when consequently adding shifted synthetic speech fragments. The results confirm that except the very beginning and end of melisma where signal is still week it is possible to employ standard LPC technique even in a case of melisma reconstruction/synthesis.

## Synthesis of melisma using artificial glottal pulse

In the following we investigate the possibility to use artificially generated signal of glottal pulses in order to apply LPC technique and to restore fragment of melisma.

Natural signal of glottal pulse possesses redundant information. In a modern system of speech signal synthesis usually it is replaced by quite simple approximations. The main reason of doing this is the need to achieve good compression ratios actual in the field of telecommunications. Worth to note that even in the synthesis based on sets of phoneme the same approach is used.

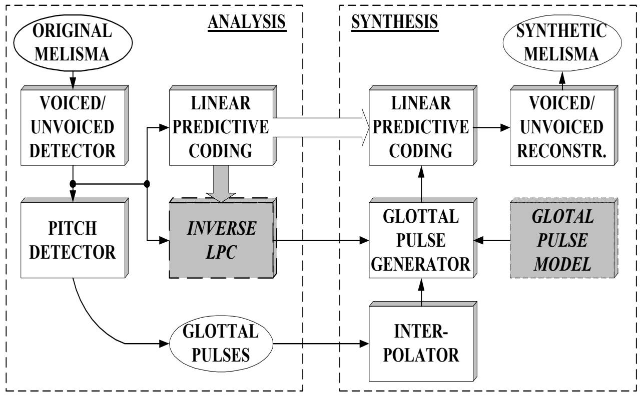


**Fig. 2.** Two cases of melisma synthesis/restoration in the one scheme. First case employs Glottal Pulse Model indicated by right grey dotted box. Second case uses Inverse LPC block shown as a grey dashed box at left

Let us examine the applicability of frequently employed glottal pulse model [12] expressed by

$$g(n) = \begin{cases} \frac{1}{2}\left[1 - \sin\left(\frac{n\pi}{P}\right)\right], & 0 \le n \le P; \\ \cos\left[\frac{(n-P)\pi}{2(K-P)}\right], & P \le n \le K; \\ 0, & n > K. \end{cases} \quad (2)$$

Here, parameters $P$ and $K$ are used to control the shape of glottal pulse.

In order to construct artificial glottal signal, first pitch signal is extracted from the fragment of melisma (see Fig. 2) then it is processed through interpolator. This let us to employ expression (2) duplicating in required time instances synthetic pulses. Result of restored *gruppett* is presented in Fig. 3a. Comparison of it with the original one (see Fig. 1a) not only visually but also perceptually during listening tests confirms, that restoration is not good — higher frequencies are missing and resultant sound possesses mechanical artifacts. Nevertheless, in the area of telecommunications such result could be sufficient.

## Synthesis of melisma using original fragment of glottal pulse

Previous results lead to conclusion, that quality of synthetic melisma is dependent on the shape of glottal pulse. Let us examine the possibility to use fragment of original glottal pulse in the synthesis of complete melisma. We will again perform synthesis according to scheme presented in Fig. 2, however in a place of Glottal Pulse Model we will use the Inverse LPC block (marked in grey) to get a fragment of original glottal pulse. Again, pitch detector and interpolator will duplicate the fragment of glottal pulse and finally using such artificial glottal signal,

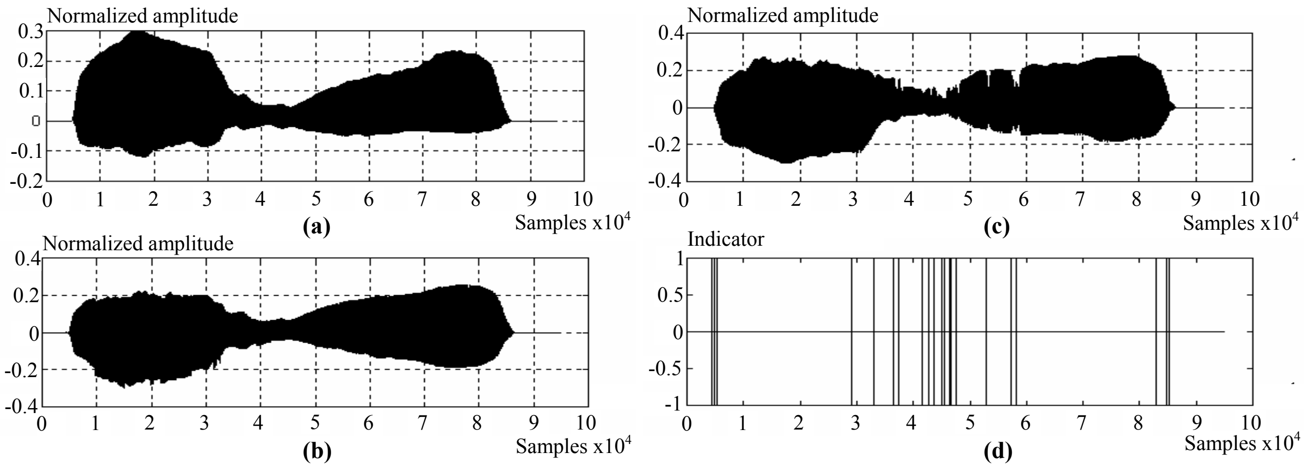


**Fig. 3.** Result of *gruppett* synthesis employing three different methods: a) synthesized *gruppett* using Glottal Pulse Model, b) synthesized *gruppett* employing single fragment of original glottal pulse, c) synthesized *gruppett* employing multiple fragments of original; glottal pulses using correlation technique, d) indicator function, showing places where glottal pulse shape was changed during synthesis presented in part (c)

melisma will be synthesized. Resulting synthetic *gruppett* is shown in Fig. 3b.

Acoustically synthetic *gruppett* does not differ from original. While testing this method on similar representatives of *gruppett* as also as on other types of melisma it was noted that results for *trill* were worst. Appeared, that timbre during *trill* varies most, while for other kinds of melisma it is not. In order to improve our method and include *trill* as possible candidate for synthesis, we introduced constant updating of original glottal pulse. Time interval of 30 ms was selected when the same glottal pulse was kept, even thought not the original, but mean glottal pulse was employed based on the calculation of cross-correlation function

$$R_{xy}^{N}(\tau) = \frac{1}{N}\sum_{t=\tau}^{N} y(n)x(n-\tau). \quad (3)$$

Here $x(n)$ and $y(n)$ are $N$ duration signals to be cross-correlated. Example of such "mean" glottal pulse fragment is shown in Fig. 4b. During experiment employment of smoothing window was not capable to remove introduced distortions stipulated by the mismatch of neighboring "mean" glottal pulses. Such obstacle was overcome slightly changing the synthesis process: we used the same "mean" glottal pulse however the variability of it was recaptured varying its amplitude. In order to determine the amount of glottal pulse adjustment, we calculate cross-correlation between neighboring signal fragments and calculate division of its maximum value with and a mean value:

$$\alpha = \frac{\max(R)}{\mathrm{mean}(R)} = \begin{cases} >10, & \text{change glottal pulse;} \\ \leq 10, & \text{use previous glottal pulse.} \end{cases} \quad (4)$$

Based on the experimentation, it was found that value 10 for the ratio is a critical one. If the ratio exceeds 10 then "mean" glottal pulse must be changed otherwise it will be updated according to the actual value of glottal pulse amplitude for this particular fragment. Results of such synthesis of *gruppett* are shown in Fig. 3c while places where adjustments of pulse were done are presented in Fig. 3d. Synthesis/restoration of melisma was not perfect — listening tests confirmed that reduced change in glottal pulses is still audible. Closer look at Fig. 1b and Fig. 3c reveals the fact, that places of glottal pulse and pitch changes coincide, leading to conclusion that there glottal pulses must be updated frequently, perhaps each 30 ms duration fragment. Another factor playing role here is amplitude of signal. When the intensity of speech signal is low, glottal pulses are buried in a noise. Our set-up naturally leads to more frequent glottal pulse update (see middle part of Fig. 3d) because of reduction of cross-correlation. Similar results were obtained in a case of *mordent*, where frequent glottal pulse update was introduced "automatically" at the beginning of melisma.

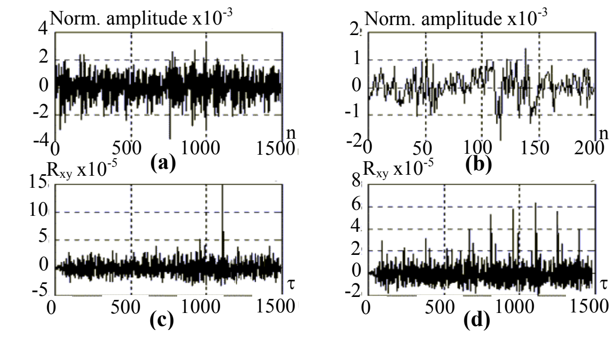


**Fig. 4.** Determination of glottal pulse using correlation method: a) 30 ms duration glottal signal, b) "mean" of 5 real glottal pulses, c) cross-correlation of glottal pulse in (b) with original one—indicates big difference; d) the same cross-correlation of (b) glottal pulse with original one—does not indicate big difference

## Conclusions

1. The problem of restoration of the comparatively long fragments of songs called melisma was proposed to be cast in a framework of a signal synthesis.
2. Several case studies of the synthesis of melisma using synthetic, single original and multiple original glottal pulses were presented.
3. Experiments confirmed that synthesis of melisma could be done using ordinary LPC method, however special attention must be called to the quality of glottal

pulses. Our study outlined that the standard static models of glottal pulse are not suitable for synthesis of melisma.

4. New method of melisma synthesis based on employment of original glottal multi-pulses and their extraction way was presented.

5. To cope with variability of pitch and intensity in a place of melisma, cross-correlation test based update of glottal pulse was suggested.


## Acknowledgements

This research work in part was supported by The Swedish Institute through New Visby programme project Ref. No. 2473/2002(381/T81) and The Royal Swedish Academy of Sciences.


## References


1. **Schuller D.** Data density versus data security: formats suitable for archival purposes // 43 FIAF Congress: Archiving the Audio-Visual Heritage. - Berlin, 1987. – P. 85-97.
2. **Navakauskas D.** Artificial Neural Network for the Restoration of Noise Distorted Songs Audio Records, Doctoral Dissertation No. 434 VGTU. –V.: Technika, 1999.
3. **Navickaitė-Martinonienė E.** Elementarioji muzikos teorija. - V.: Vaga. –1979.
4. **Žintelytė M.** Review of Speech Synthesis Methods // Information Technology'99: Conference Proceedings. – Kaunas: Technologija, 1999. – P. 226-230.
5. **ScanSoft**. RealSpeak™ System // http://www.lhsl.com.
6. **Sutton S., Novick D. G., Cole R. A., Fanty M.** Building 10,000 spoken-dialogue systems // Proceedings of the International Conference on Spoken Language Processing. – Philadelphia, PA, 1996.
7. **Sproat R.** Multilingual Text-to-Speech Synthesis: The Bell Labs Approach, Kluwer Academic Publishers. - Boston, 1997.
8. **Black A. W., Taylor P. A.** CHATR: A generic speech synthesis system, COLING'94, Kyoto, Japan, 1994. – P. 4
9. **Telcordia**. Hybrid ORATOR® II Speech Synthesizer // http://www.argreenhouse.
10. **Elan Speech**. Elan Text to Speech™ and ProVerbe Speech Unit™ // http://www.elanTTS.com.
11. **Collins** English Dictionary. Updated Edition, 1994.
12. **Deller J. R., Proakis J. G., Housen J. H. L.** Discrete-Time Processing of Speech Signals. - New York: Macmillan, 1993.



Pateikta spaudai 2003 05 19



**R. Leonavičius, D. Navakauskas. Melizmų sintezės aspektai // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2003. – Nr. 6(48). – P. 18-21.**

Melizmų restauravimo uždavinį siūloma spręsti signalų sintezės metodais. Nagrinėjamos galimybės sintezuoti melizmas naudojant originalius, aproksimuotus bei dirbtinius melizmų parametrus. Tiriami sintezės metodai: naudojant tiesinio prognozavimo koeficientus ir tikrą gerklų signalą bei tiesinio prognozavimo koeficientus, dirbtinį gerklų signalą, stiprinimą ir pagrindinį toną. Siūloma ir nagrinėjama melizmų sintezės metodika naudojant originalaus gerklų signalo fragmentą bei originalius tiesinio prognozavimo koeficientus, stiprinimą ir pagrindinį toną. Kalbos signalų charakteristikų kitimui įvertinti siūloma naudotis paprasta, tačiau veiksminga kryžminės koreliacijos skaičiavimo procedūra. Il. 4, bibl. 12 (anglų kalba; santraukos lietuvių, anglų ir rusų k.).

**R. Leonavičius, D. Navakauskas. Aspects of Melisma Synthesis // Electronics and Electrical Engineering. – Kaunas: Technologija, 2003. – No. 6(48). – P. 18-21.**

The problem of restoration of the comparatively long fragments of songs called melisma was proposed to be cast in a framework of a signal synthesis. The paper investigates several possibilities to synthesize melisma employing original, approximated and modeled parameters of melisma. Several synthesis ways are explored: employing linear prediction coding (LPC) coefficients and original glottal signal, and LPC coefficients, modeled glottal signal, gain and pitch. New method for synthesis of melisma employing fragment of original glottal signal and original LPC coefficients, gain and pitch is proposed. In order to cope with the variability of speech signals simple but effective procedure based on the cross-correlation test is proposed. Ill. 4, bibl. 12 (in English; summaries in Lithuanian, English, Russian).

**Р. Леонавичюс, Д. Навакаускас. Аспекты синтеза мелизм // Электроника и электротехника. – Каунас: Технология, 2003. – №. 6(48). – С. 18-21.**

Задачу реставрации мелизма (фрагмента песни с сравнительно долгой продолжительностью) предполагается решать используя теорию синтеза сигналов. Исследуется возможность использования для этих целей оригинальных и синтетических параметров мелизма. Несколько методов синтеза сигналов исследуется с целью применения для реставрации мелизма. Предлагается методика синтеза мелизма с использованием коэффициентов линейного прогноза и фрагмента голосового импульса. Для того, чтобы учесть изменения характеристик мелизма, предлагается использовать простую и эффективную процедуру на базе исчисления корреляции. Ил. 4, библ. 12 (на английском языке; рефераты на литовском, английском и русском яз.).