# Amplitude and Frequency Modulation in Speaker Recognition Systems

Zygmunt Ciota, *Member, IEEE*

***Abstract*—The paper presents a review of the nowadays methods of voice vector extraction, applied in such speech processing, like person identification and emotion recognition. A special attention was held on mixed time-frequency analysis based on temporary frequency approach. The methods of calculation of time – frequency voice characterization were also described. The most important building blocks of identification and recognition of speakers have been presented. The characterization of feature vectors suitable for identification and verification in microcomputer systems was described. Components and appropriate method of speech identification based on the long-term spectra vectors were discussed.**

***Index Terms*—AM-FM modulation, Gabor filters, Hilbert transform, Speaker recognition system, Spectrogram analysis, Speech processing.**

## I. Introduction

ADVANCED computer systems of person identification, based on voice analysis, characterize a similar precision in comparison with an natural identification provided by a man. Currently, an effort of researches is focused on an improvement of a big human population identification, going over a natural recognition capability of a single person. Unfortunately, we still have not the precise understanding of discriminatory mechanisms and moreover, well defined feature vectors are not available.

The idea of such researches is going from sixty years of XX century, when L.G. Kersta suggested in his paper [1], that spectrogram analysis should be sufficient to person identification in a similar way, like fingerprint analysis. Currently it is impossible to collect a set of voice parameters in a similar way like a minutiae-based fingerprint verification system.

Person recognition can be proceeded by using two different ways: identification or verification. In the first case you are working with a close confined system, under assumption that the analyzed voice belongs to the data base and it will be necessary to find a proper person using classification approach. A verification process takes place in an open system, it means you are checking if the voice belongs or not to the data base of our privileged persons.

Z. Ciota is with the Department of Microelectronics and Computer Science, Technical University of Lodz, Wolczanska 221/223, 90-924 Lodz, Poland, (e-mail: ciota@dmcs.pl)

## II. Time Frequency Representation of the Speech

One of the most important stage of recognition process is parameter extraction of speech signal. Methods based on spectrogram analysis offer several time insufficient accuracy, therefore in the ninety years of XX century, a new method has been elaborated. The method bases on another time–frequency diagram analysis, and is called often as the speech pyknogram [2, 3, 4], characterizing time varying resonances of a vocal tract. The proposed approach permits to extract more precisely the frequencies, responsible for a formulation of the phoneme formants. In the other words, you can extract with higher precision such speech features, like formant tracking, calculation of Mel-Frequency Cepstral Coefficients (MFCCs). Afterwards, the time varying properties of voice behavior and calculation of low-frequency resonances can be established in a better way.

Standard analysis of a voice signal takes into account a vocal tract model based on an input acoustic signal generated in a glottis and passing through time-varying filter as a tract representation. In such approach a several occurrences have been neglected. The most important are: an influence of non-stabilized air outflow from the lungs causing activation of the glottis, air turbulences in the vocal tract, oscillations of some components of the tract, and also fluctuations of muscles controlling the speech process.

Among several attempts of a precise characterization of speech process phenomena, the methods based on pyknogram developments, seems to be very efficient and they are widely applied in nowadays voice analysis. In particularly, a model delivered form AM-FM (Amplitude Modulation – Frequency Modulation) approach is very promising. AM-FM model is based on speech signal decomposition into non-correlated channels of bandpass. Each channel is characterized by the amplitude envelope and the phase, creating the temporary frequencies.

To find parameters of a real voice signal $u(t)$, represented usually as microphone output voltage, it will be necessary to build analytical signal $u_a(t)$, according to the following expression [2]:

$$u_a(t) = u(t) + j \cdot \hat{u}(t) \tag{1}$$

where $\hat{u}(t)$ is a Hilbert transform of the signal $u(t)$.

Analytical signal can be expressed in the following manner:

$$u_a(t) = a(t) \cdot e^{j\varphi(t)} \tag{2}$$

where $a(t)$ is a temporary amplitude of analytical signal and $\varphi(t)$ represents the phase.

Temporary frequency $f(t)$ of analytical signal $u_a(t)$, can be find as the phase differential:

$$f(t) = \frac{1}{2\pi} \cdot \frac{d\varphi(t)}{dt} \tag{3}$$

Building process of the pyknogram consists of calculation and graphical presentation of temporary resonant frequencies. Significant influence of such frequencies results from non-stationery nature of speech signal. Therefore, they can be applied in person identification and emotion recognition methods.

Such representation of a speech concerning the same expression repeated several time by the same person, can be different in a similar way like in the case of spectrograms. It results principally from the randomly changers observed in a speech generation. Furthermore, it can be also an effect of emotional state of the speaker or health condition. As an example, one can notice, that a simple cold causes significant changes in nasal tract characteristics. On the other side, a man can change his voice intentionally. Decreasing the slot of the glottis one can increase the fundamental frequency, it is also possible to modify a geometry of a vocal tract, modeling intentionally a floppy components of the tract, like language or palate.

Between spectrogram and pyknogram one can observe a significant difference. In the first case you have a three dimensional diagram, where a surface is represented by two axes representing time and frequency, while the color or a darkness level corresponds to the intensity of a given frequency in a given time while. The pyknogram is also represent in a surface confined by time and frequency axes, but a diagram is illustrated by a set of points of temporary channel resonances. Therefore, in the case of similar formant frequencies (or even overlapping resonances) that area becomes less readable. Moreover, such resonant frequencies can be calculated by using different filtering systems, therefore the pyknograms of the same expression, spoken by the same person, may be significantly different.

## III. Spectrogram and Pyknogram Representations

The pyknogram can be find using different filtering systems well known from the theory of signal processing dedicated to acoustic frequency signals [5]. Currently, one can observe in a literature new methods and especially modifications of already existing filtering approaches [6, 7].

An important algorithm of AM-FM model calculation will be presented in this chapter. Upon the presented approach, one can be evaluated the time and memory resources necessary to effective application to voice identification and speaker recognition. As the first step it will be necessary to apply Multiband Demodulation Analysis (MDA) [8, 9]. Afterwards, every filtering bandpass should be demodulated to the instantaneous amplitude and frequency. As a filter bank, predominatingly Gabor filter set will be applied. The central passband frequencies should be located evenly along the entire acoustic range. Gabor filters are characterized by a smooth time and frequency characteristics and are very useful in precise calculations of temporary amplitude and temporary frequency in demodulation process.

A speech signal $u(t)$, after filtering process realized by using passband filter bank, is decomposed for a set of signals $w_i(t)$, where $i$ denotes an output of $i$-th filter. Afterwards, for every $w_i(t)$, Hilbert transform $\hat{w}_i(t)$ has been calculated. The instantaneous amplitude of the signal $a_i(t)$ for every of bandpass is calculated as follows:

$$a_i(t) = \sqrt{\hat{w}_i^2(t) + w_i^2(t)} \tag{4}$$

Instantaneous frequency of a speech signal for $i$-th bandpass can be obtained by differentiating of the phase $\varphi_i(t)$:

$$f_i(t) = \frac{1}{2\pi} \cdot \frac{d\varphi_i(t)}{dt} = \frac{1}{2\pi} \cdot \frac{d}{dt}\left\{\arctan\left[\frac{\hat{w}_i(t)}{w_i(t)}\right]\right\} \tag{5}$$

Based on an instantaneous amplitude and frequency one can obtain a short-time estimate $F_i$ of weighted-average instantaneous frequency for every $w_i(t)$:

$$F_i(t) = \frac{\int_{t_0}^{t_0+\tau} f_i(t) a_i^2(t) dt}{\int_{t_0}^{t_0+\tau} a_i^2(t) dt} \tag{6}$$

where $\tau$ is the length of the time frame and can be changed by the user. In a similar way like in the case of spectrograms, the overlapping bands $w_i(t)$ should be applied.

A very important task is a proper choice of passbands in the case of Gabor filters, premising for pyknogram a control in such a way, that you can obtain the maximal enhancement of characteristically important speech features, effectual for a given current application. In a similar way like in other methods, for example: GMM, LPC, MFCC, the most often it will be formant identification and tracking, and also a searching of quasi-harmonic vibrations, responsible for the generation of glottis harmonic components.

Parameters of feature components for the necessity of voice identification are mostly represented by different frequency

values and are organized as matrices at dimensions $G$ x $N$, where $G$ is the number of Gabor filters and $N$ denotes time frames $\tau$ calculated for the entire analyzed utterance. For every $\tau$ the corresponding values are estimated as instantaneous frequencies according to the analysis of Gabor filter outputs.

Several methods based on pyknogram analysis give very good results in identification and verification systems, being better in comparison with the systems based on extraction of cepstrum coefficient parameters or Gaussian mixed models [10, 11, 12].

## IV. Designing of Identification System Components

The most important component of identification system is a base of feature vectors of known persons. The base is created during training and teaching processes and will be exploited during identification of an unknown person. One of the important task is a proper definition of parameters, building a feature vector. According to the established recognition precision and a largeness of identified population, the system can demand different resources of computer memory and different clock frequency of the processor.

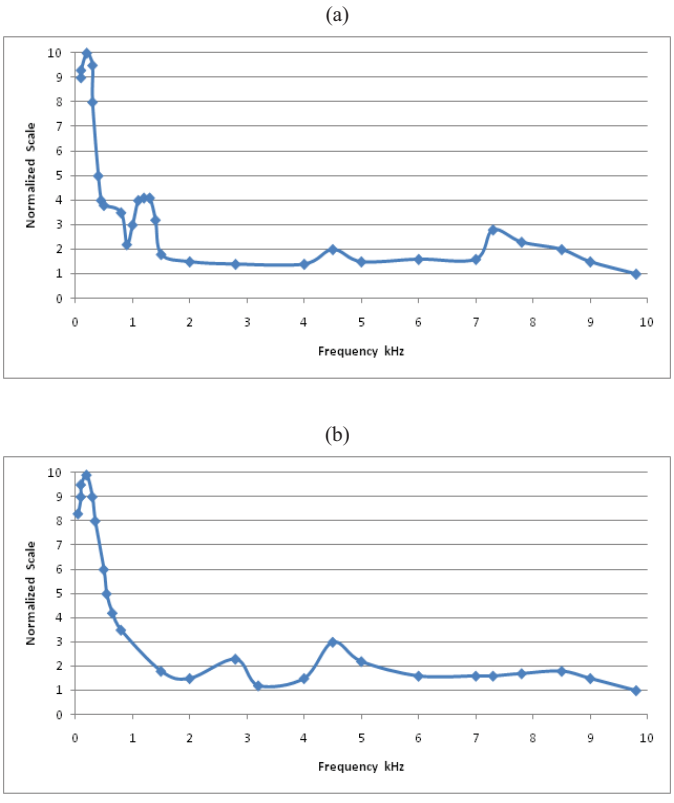


Fig.1. Long-term spectra vectors presented in the diagram forms. Two vectors (a) and (b) have been obtained for two different male persons, according to the 8 sec. length utterances. The $y$ axis is normalized to the 10 point scale.

In the case of the population consisting of one – two hundreds of people, resources of a nowadays microcomputer will be sufficient. In such a case several feature parameters can be chosen. As an example one can mention the following group of parameters: Long-Term Spectra Vector (LTS), Speaking Fundamental Frequency (SFF), Time-Energy Distribution (TED), and Vowel Formant-Tracking (VFT) [13].

In the case of identification procedure, one of the most important features can be stored in LTS vector, and it will be calculated using the average long term attribute. In this case the energy of speech spectrum should be chosen, therefore the information of a tone and a power distribution of the voice are included.

According to the laboratory experiences, the range over 10 kHz contains meaningless information, on the other hand, the telephone bandpass of only 4 kHz, is insufficient in commercial application for a medium population of identifying persons. LTS vector can be obtained using different methods, for example Fourier transform, linear coding of prediction, using cepstrum coefficients, etc. Fig. 1 presents examples of two LTS vectors in a diagram form for two different speakers.

To increase a precision of the system it is possible to increase of parameter number applying alternating vectors. It is important to know, that SSF vector carrying information concerning fundamental glottis frequency, TED is sensitive to the prosody of the voice, and VFT is in a similar way like LTS, insensitive to the voice distortion.

The TED vector can be represented by several parameters chosen according to the requirements of the identification system. The most important parameters of TED vector are presented below:

- total time of registered speech, defined as a sum of all syllables of the utterance,
- number of silence intervals,
- ratio of speech time to the total silence,
- ratio of total time of registered speech to the time of entire utterance,
- coefficient of speech time, representing effectiveness of speech processes,
- speech ratio, representing a number of single syllable speaking in the unity time.

After a proper choice of the set of feature parameters, the identification system can be competed using the following components:

- block of registration and playback of a speech signal,
- signal normalization,
- spectrogram or pyknogram calculation,
- feature vector calculation and normalization,
- teaching block and pattern of references definition,
- calculation of different vectors distances and verification.

As a distance between different vectors several method can be applied, the simplest and still very useful are nearest neighbor and nearest mean sorters.

## V. Conclusion

A problems occurring during speech recognition are significantly more difficult in comparison with speaker identification. In both cases it is necessary to use another groups of feature vectors. In the case of speech recognition, the structure of formant should remain the same for different speakers, therefore formant analysis is very important. On the other hand, identification systems demand vectors, which are sensitive to intonation, emotion and particularly, you need high sensitive parameters to find a proper glottis frequency characteristics.

Identification process must be often provided during telephone or informal conversation, therefore one cannot rely on the same predefined utterance. As a consequence, the long-term parameters are usually preferred in identification systems. Nevertheless, the designers of such systems have to solve another difficulties, resulting from unexpected changes of voice characteristic, caused by unexpected health state and speech variability as a unknown function of time, counted sometimes in year units [14, 15].

## References


[1] L. G. Kersta, "Voiceprint identification", *Nature*, vol. 196, pp. 1253–1257, 1962

[2] Saeed Gazor, Reza Rashidi Far, Adaptive Maximum Windowed Likelihood Multicomponent AM-FM Signal Decomposition, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, March 2006

[3] M. H. Hayes, J. S. Lim & A. V. Oppenheim, "Signal Reconstruction from Phase or Magnitude", *IEEE Trans. Acous. Speech & Signal Proc*., vol. ASSP-28, No. 6, pp. 672-680

[4] Yasser Hifny, Steve Renals, Speech Recognition Using Augmented Conditional Random Fields, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, February 2009

[5] T. P. Zieliński, Cyfrowe przetwarzanie sygnałów. Od teorii do zastosowań, *Wydawnictwa Komunikacji i Łączności*, Warszawa 2009

[6] Thomas Pellegrini, Lori Lamel, Automatic Word Decompounding for ASR in a Morphologically Rich Language: Application to Amharic, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, July 2009

[7] D. V. Dimitriadis, P. Maragos, A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Process. Letters*, vol. 12, no. 9, Sepember 2005

[8] Ran D. Zilca, Brian Kingsbury, J. Navrátil, Ganesh N. Ramaswamy, Pseudo Pitch Synchronous Analysis of Speech With Applications to Speaker Recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, March 2006

[9] Mari Ostendorf et al., Speech Segmentation and Spoken Document Processing, IEEE Signal Processing Magazine, May 2008

[10] Z. Ciota: „Metody przetwarzania sygnałów akustycznych w komputerowej analizie mowy". *Akademicka Oficyna Wydawnicza EXIT*, Warszawa 2010

[11] Marco Grimaldi, Fred Cummins, Speaker Identification Using Instantaneous Frequencies, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, August 2008

[12] S. Grocholewski, Two Stage Speaker Verification System, *Speech and Language Technology*, vol.6, 2005, pp.45-56

[13] Z. Ciota: "Audio-Haptic Feedback in Speech Processing". *The 6th IEEE Int. Workshop on Haptic, Audio and Visual Environments and Games* - HAVE 2007, 12-14 October 2007, Ottawa, Canada

[14] T. R. van de Water, Otolaryngology: Basic Science and Clinical Review, Stuttgart, Thieme Publishing Group, 2005

[15] J. C. Stemple, L. E. Glaze & B. Klaben Gerdemann, Clinical Voice Pathology Theory and Management, 3rd Edition, New Jersey, Thomson Delmar Learning, 2000


**Zygmunt Ciota** was born in Gryfice, Poland, on February 21, 1949. He received the M.Sc., Ph.D. and D.Sc. degrees from Technical University of Lodz in 1973, 1984 and 1996 respectively. Since 1973 until 1979 he was employed in the industrial enterprises. Since 1979 until now he is with the Technical University of Lodz, working at the Institute of Electronics until December 1996, and next he joined the Department of Microelectronics and Computer Science. He is the author or co-author of over 100 scientific publications. Z. Ciota was the head of 7 grants of the Polish Committee of Scientific Research, and he was also the participant of 11 international projects, concerning education and research. He is interested in VLSI design of mixed digital-analog systems, microsystems, computer-aided modeling of semiconductor devices and signal processing.