УДК 004.4'277

SPECTRUM ESTIMATION OF SPEECH: CODING AND FEATURE EXTRACTION

M. Taha
PhD Student

I. Azarov
Doctor of Technical Sciences

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: mostophataha@hotmail.com, azarov@bsuir.by

M. Taha

14.04.63, graduated from Lebanese University (1986), specialization – computer and telecommunication. Master of science. Lecturer at Arts, Sciences and Technology University in Lebanon, Jadra. Professional interests: speech processing.

I. Azarov

11.07.1980, graduated from Belarusian State University (2002) as mathematician. Holds doctorate degrees (Candidate of Science, 2009; Doctor of Science, 2015) from Belarussian State University of Informatics and Radioelectronics. Specialization – informatics. Professional interests: digital signal processing.

Abstract. Speech analysis and spectrum estimation has been the fundamental problem of digital signal processing for recent decades. The problem still has a huge practical impact on modern speech processing applications that involve coding and deep learning. The paper reviews the main speech spectral estimation techniques including linear prediction and cepstrum.

Keywords: linear prediction, speech coding, spectrum estimation

1. Introduction. All the speech sounds in the language can be grouped into two groups: vowels and consonants. The main difference between vowels and consonants lies in their articulation; a vowel is articulated with an open vocal tract whereas a consonant is articulated with complete or partial closure of the vocal tract.

1.1. Vowels. In vowels, the tongue does not touch the teeth, lips or the roof of the mouth when pronouncing vowel. It is important to notice that the term vowel can refer to both vowel letters and vowel sounds in the alphabet. There are five vowel letters in the alphabets: a, e, i, o, and u. Also, there are about twenty vowel sounds in English language, and these five vowel letters represent all these sounds. This means that a single vowel letter can represent many vowel sounds. For example, consider the words cat, heart, hate, all, and said. All these words contain the vowel a, but the sound of the vowel is different in each word. In the same manner, the same vowel sound can be also represented by different vowel letters as well. For example, note how the sound /ei/ is represented by different words in the words weigh, clay, bake, and opaque. All vowels in English

are voiced. That means the vocal cords vibrate while we say a vowel. To feel this voicing, touch your throat and say AAAA. Feel that movement in your neck.

1.2. Consonant. Since the vocal tract is not completely open, the sound is stopped by teeth, tongue, lips, or constriction of the vocal cords at a certain point. The term consonant also refers to both letters and sounds. All the letters in the alphabet, apart from the 5 vowels, are consonants. Thus, there are 21 consonant letters in the alphabet, and they represent about 24 consonant sounds. Consonants can be grouped into two major categories: voiced and unvoiced consonants. Voiced consonants are obtained when the vocal cords vibrate; whereas unvoiced consonants don't. It is easy to identify voiced and unvoiced consonants by putting your finger on the throat and feeling the vibrations. [1]

2. Speech Coding Techniques

2.1 Definition. Speech coding is the process of transformation the speech signal in a more compressed form, which can be transmitted or stored with few numbers of binary digits. Speech coding techniques are mainly two types: Lossless (the reconstructed signal is identical to the original one) and lossy (reconstructed signal is different from the original one).

In general, speech coding is a lossy type of coding. Because, it removes the information which are irrelevant from the perceptual quality point of view.

2.2 Codec. It's a device or computer program capable of encoding or decoding a digital data stream. In software, it's a computer program implementing an algorithm that compresses and decompresses digital audio data. The objective is to represent the high-fidelity audio signal with minimum number of bits while retaining quality. In hardware, audio codec refers to a single device that encodes analog audio as digital signals and decodes digital back into analog.

Many low bit-rate speech coders represent the spectral envelope by an all-pole digital filter whose coefficients are calculated by a form of linear prediction (LP) analysis. The lower the bit-rate, the more critical will be the accuracy of the spectral analysis for achieving good quality speech. Here are some techniques of speech coding:

2.2.1 MELP. Mixed Excitation Linear Prediction (MELP). It's used mainly in military applications, satellite communications, secure voice, and secure radio devices. The MELP algorithm specifies the 2400 bps mode of operation with an optional post filter available for the decoder. The MELP frame size is 22.5 ms in duration with each frame containing 180 digital voice samples at a sampling rate of 8,000 Hz. [2]

2.2.2 OPUS. Opus combines the speech-oriented linear predictive coding SILK[1] algorithm and the lower-latency, MDCT[2] algorithm. It switches between or combine them as needed for maximal efficiency. Bitrate, audio bandwidth, complexity, and algorithm can all be adjusted smoothly in each frame. Opus can handle a wide range of audio applications, including Voice over IP, videoconferencing, in-game chat, and even remote live music performances. It can scale from low bitrate narrowband speech to very high quality stereo music. Supported features are:

- Bitrates from 6 kb/s to 510 kb/s
- Sampling rates from 8 kHz (narrowband) to 48 kHz (fullband)
- Frame sizes from 2.5 ms to 60 ms
- Support for both constant bitrate (CBR) and variable bitrate (VBR). [3]

2.2.3 GSM. Global System for Mobile communications is a cellular phone system standard. It includes a codec[3]. It uses the information from previous samples (this information does not change very quickly) in order to predict the current sample. The speech signal is divided into blocks of 20 ms. These blocks are then passed to the speech codec, which has a rate of 13 kbps, in order to obtain blocks of 260 bits. Newer GSM systems use a couple of newer codecs. However these are heavily patent encumbered:

[1] SILK is an audio compression format and audio codec developed by Skype Limited.

[2] Modified Discrete Cosine Transform.

[3] Device or program that compresses data to enable faster transmission and decompresses received data.

- EFR (Enhanced Full Rate) uses ACELP (Algebraic Code Excited Linear Prediction)
- HR (Half Rate) uses CELP (Code Excited Linear Prediction). [4]

3 Linear Predictive Coding (LPC). It's a tool used mostly in audio signal processing and speech processing for representing the spectral envelope (the envelope curve of the amplitude spectrum) of a digital signal of speech in compressed form, using the information of a linear predictive model. It is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters. [5]

3.1 Envelope Calculation. The LPC method is quite close to the FFT (Fast Fourier Transform)[4]. The envelope is calculated from a number of formants or poles calculated through an equation.

- The formants are estimated removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The removing process is called inverse filtering, and the remaining signal is called the residue (prediction error).
- The speech signal – source – is synthesized from the buzz parameters and the residue. The source is ran through the filter – formants –, resulting in speech.
- The process is iterated several time is a second, with "frames". A 30 to 50 frames rate per second yields and intelligible speech. [6]

3.2 Linear Prediction. A speech signal can be written in the following form:

$$x(k) = \sum_{l=1}^{L} a_l \, x(k-l) + Gu(k)$$

where k is the time index, L represents the number of coefficients in the model (the order of the predictor), a_l, l=1, 2, …., L are defined as the linear prediction coefficients, G is the gain of the system, and u(k) is the excitation signal which can be either a quasiperidoc train of impulses or a random noise source. The periodic source produces voiced sounds such as vowels and nasals, and the noise source produces unvoiced sounds such as the fricatives.

The previous equation can be rewritten in the frequency domain by using the z-transform. If H(z) is the transfer function of the system, we have:

$$H(z) = \frac{G}{1-\sum_{l=1}^{L} a_l z^{-l}} = \frac{G}{A(z)}$$

which is all poles transfer function. This filter H(z) is a good model of the human vocal tract. Our main concern is to determine the predictor coefficients a_l, l=1, 2, …. L, and to study the properties of the filter A(z).

3.3 Forward Linear Prediction. Consider a stationary random signal x(k). The objective of the forward linear prediction is to predict the value of the sample x(k) from its past values, i.e., x(k-1), x(k-2), etc. we define the forward prediction error as : $e_{f,L(k)} =$ x(k)– $\hat{x}$(k) = $x(k) - \sum_{l=1}^{L} a_{L,l} x(k-l) = x(k) - a_L^T \, x(k-1)$. Where the superscript 'T' denotes transposition, $\hat{x}$(k) is the predicted sample. $a_L = \left[a_{L,1} \; a_{L,2} \; \dots\dots\dots a_{L,L}\right]^T$ is the forward predictor of length L, and $x(k-1) = [x(k-1) \;\; x(k-2) \;\; \dots\dots\dots x(k-L)]^T$ is a vector containing the L most recent samples starting with and including x(k-1).

[4] FFT converts a signal into individual spectral components and thereby provides frequency information about the signal. FFT is an optimized algorithm for the implementation of the "Discrete Fourier Transformation" (DFT). A signal is sampled over a period of time and divided into its frequency components. These components are single sinusoidal oscillations at distinct frequencies each with their own amplitude and phase.

We would like to find the optimal Wiener predictor. For that, we seek to minimize the mean-square error (MSE): $J_f(a_L) = E\{e_{f,L}^2(k)\}$ were E{.} denotes mathematical expectation. Taking the gradient of $J_F(a_L)$ with respect to a_L and equating to 0_{Lx1} (a vector of length containing only zeros), we easily find the Wiener-Hopf equations: $R_L a_{o,L} = r_{f,L}$ where the subscript 'o' in $a_{o,L}$ stands for optimal and

$$R_L = E\{x(k-1)x^T(k-1)\} = E\{x(k)x^T(k)\}$$
$$= \begin{pmatrix} r(0) & r(1)\cdots & r(L-1) \\ r(1) & r(0)\ldots & r(L-2) \\ \vdots & \vdots & \vdots \\ r(L-1) & r(L-2)\cdots & r(0) \end{pmatrix}$$

is the correlation matrix, and $r_{f,L} = E\{x(k-1)x(k)\}$ is the correlation vector. The matrix R_L has a Toeplitz structure (i.e., all the entries along the diagonals are the same); assuming that it is nonsingular, we deduce the optimal forward predictor: $a_{o,L} = R_L^{-1} r_{f,L}$ [7, 8].

There are two main kinds of predictor estimation methods: autocorrelation, that provides a stable filter and covariance that provides a more detailed spectrum estimate. Four techniques of spectrum estimation are shown in figure 1: short Fourier transform, LP modified covariance method (20 coefficients), LP modified covariance method (128 coefficients), LP autocorrelation method (20 coefficients).

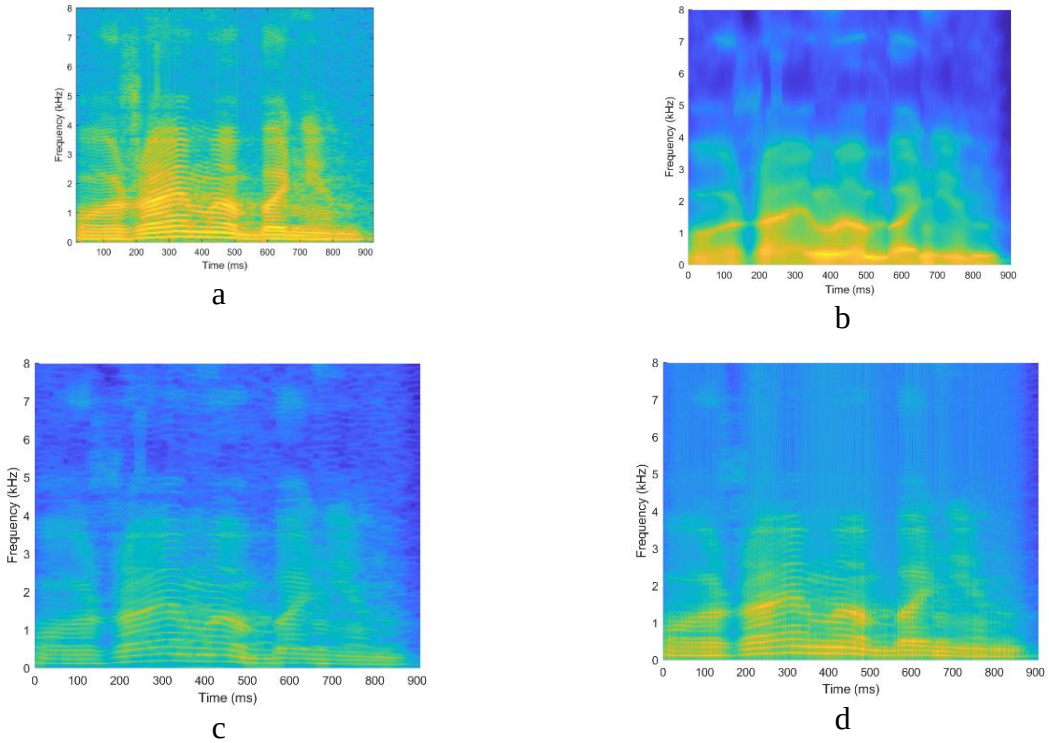

Figure 1. Amplitude spectrum estimation of a speech sample
a – short Fourier transform,
b – linear prediction, modified covariance method (20 coefficients),
c – linear prediction, modified covariance method (128 coefficients),
d – linear prediction, autocorrelation method (20 coefficients)

4. Cepstrum

Cepstrum is a technique of parametrical spectral representation that widely used today for feature extraction for machine learning problems.

4.1 Spectrum. A sound spectrum displays the different frequencies present in a sound. It is usually presented as a graph of either power or pressure as a function of frequency. The power or pressure is usually measured in decibels and the frequency is measured in vibrations per second (or Hz). You can think of the sound spectrum as a sound recipe: take this amount of that frequency, add this amount of that frequency etc until you have put together the whole, complicated sound.

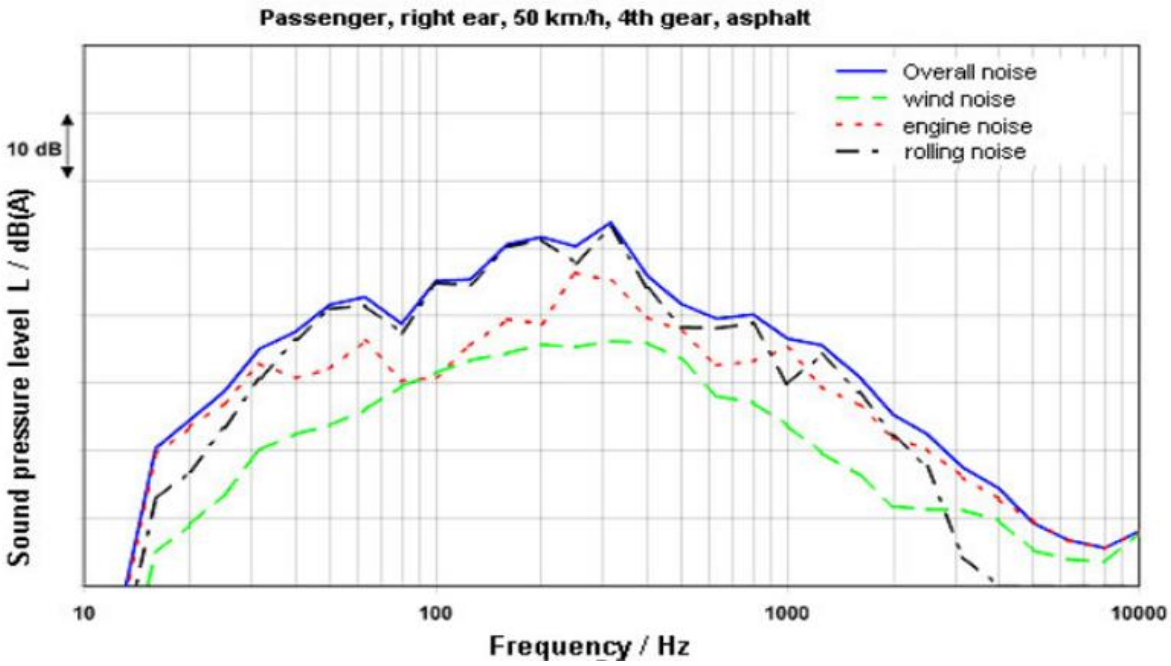

Figure 2. overall noise and noise components heard by a passenger in a vehicle at v=50km/h

4.2 Cepstrum. A cepstrum is the result of taking the inverse Fourier Transform (IFT) of the logarithm of the estimated spectrum of a signal. The name "cepstrum" was derived by reversing the first four letters of "spectrum". Operations on cepstra (many cepstrum) are: quefrency analysis, liftering, or cepstral analysis. The power cepstrum (one kind of cepstrum) finds applications in the analysis of human speech. It's an efficient tool for finding different harmonic families in the Power Spectral Density (PSD) of a time series. A power cepstrum is the inverse FFT transform of the natural logarithm of the PSD. You can compute the power cepstrum of a time series as follows: $C(\tau) = FFT^{-1}(\log(PSD))$. The power cepstrum $C(\tau)$ is a real-valued time series.

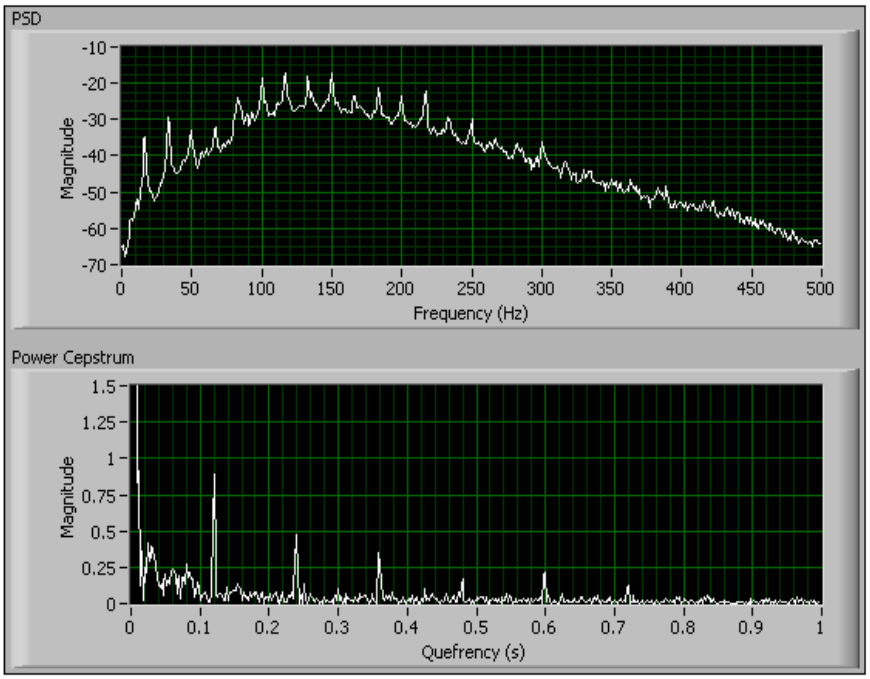

Figure 3. PSD and power cepstrum of a gearbox vibration signal

The PSD graph suggests that the signal contains both periodic and non-periodic components. In the Power Cepstrum graph, you can see that the power cepstrum gives a more clear indication of harmonic peak families than the PSD. You can modify, or lifter, the power cepstrum and then transform it back to the PSD. The word *lifter* is an anagram of the word *filter*, formed by reversing the first three letters. By liftering the unnecessary harmonic peak families, you can remove an individual peak family from the PSD.

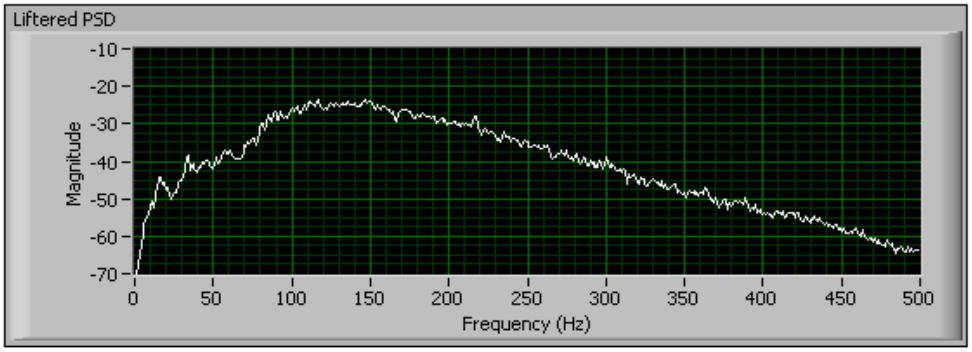

Figure 4. Liftered PSD of the gearbox vibration signal
after liftering all the harmonic peak families.

The harmonic peaks in the liftered PSD have disappeared.

To lifter the harmonic families, complete the following steps:

- Compute the power cepstrum of the vibration signal from the PSD.
- Remove harmonic peaks of individual harmonic family that you want to discard from the power cepstrum.
- Reconstruct the PSD from the liftered power cepstrum.

If a PSD contains several harmonic families, use the TSA Lifter PSD VI to separate harmonic peaks in the PSD by computing the power cepstrum and filtering the unnecessary harmonic peaks in the cepstrum [9].

References

[1] http://pediaa.com/difference-between-vowels-and-consonants/. – Date of access: 8.02.2019.

[2] https://www.vocal.com/speech-coders/melp/. – Date of access: 8.02.2019.

[3] "Opus Codec". Opus (Home page). Xiph.org Foundation. Retrieved July 31, 2012. – Date of access: 8.02.2019.

[4] https://www.voip-info.org/gsm-codec/. – Date of access: 8.02.2019.

[5] Beigi, Homayoon (2011). Fundamentals of Speaker Recognition. Berlin: Springer-Verlag.

[6] http://support.ircam.fr/docs/AudioSculpt/3.0/co/LPC.html. – Date of access: 8.02.2019.

[7] Bellangen Adaptive Digital Filters and Signal Analysis, Marcel Dekker, New York, 1987. – Date of access: 8.02.2019.

[8] S. Haykin: Adaptive Filter Theory, 4th edition, Prentice Hill, upper saddle river 2002

[9] http://zone.ni.com/reference/en-XX/help/371419D-01/lvasptconcepts/tsa_power_cepstrum/. – Date of access: 8.02.2019.

ОЦЕНКА СПЕКТРА РЕЧИ: КОДИРОВАНИЕ И ВЫДЕЛЕНИЕ ОСОБЕННОСТЕЙ

М. Таха
Доктарант

И. Азаров
Доктор технических наук

Белорусский государственный университет информатики и радиоэлектроники,
Республика Беларусь
E-mail: mostophataha@hotmail.com, azarov@bsuir.by

Аннотация. Анализ и оценка кратковременного спектра речи остается фундаментальной проблемой на протяжении последних десятилетий. В настоящее время эта проблема имеет большой практическое значение в современных приложениях обработки речи включая вокодеры и системы на основе глубокого обучения. В настоящей работе рассматриваются основные способы оценки спектра включая линейное предсказание и кепстр.

Ключевые слова: линейное предсказание, кодирование речи, оценка спектра