Perceptually Motivated Non-Intrusive Assessment of Speech Quality

Christophe Veaux and Vincent Barriac
France Telecom R&D, DIH/EQS
2, ave Pierre Marzin, 22307 Lannion, France
qualitest.ball003@rd.francetelecom.com

1. Abstract

In this contribution, we present a non-intrusive assessment technique that aims to reproduce some human perceptual mechanisms. We model a set of perceptually relevant features extracted after an auditory transform of clean speech. This internal source-model is used to characterize any occurring distortions. Preliminary results are given to illustrate the characterization of distortions by some perceptual features. It seems that the temporal resolution of our auditory transform provides interesting information to characterize non-linear distortions induced by low bit-rate coders.

2. Introduction

With the ITU-T recommendation P.862 [1], accurate predictions of perceived telephony speech quality are now achievable in an intrusive approach. Such performances rely on advanced perceptual and cognitive modeling, which allow dealing with a wide range of distortion types. However, intrusive methods require the withdrawal of connection under test whereas it would be preferable to monitor continuously the quality of speech delivered to costumers. For such case, a non-intrusive method is attractive, utilizing the in-service signal to make predictions of quality. Current non-intrusive assessment equipment performs measurements such as echo, delay, noise and loudness [2] or attempt to predict the clarity of a connection [3]. However, such simple measures cannot deal with distortions introduced in the speech structure itself by low bit-rate speech coders. Thus, non-intrusive assessment of speech quality is still an emerging field.

A non-intrusive assessment technique must address two major issues. The first is to estimate occurring distortions, it is a significant challenge of non-intrusive scheme since the original speech signal is not known. The second issue is to predict the subjective impact of the estimated distortions.

A first approach of non-intrusive assessment is to detect a set of well-characterized distortions, like "impulsive noise" or "robot voice" [4], and to learn a statistical relationship between this finite set and subjective opinions. In this case, the lack of signal reference is compensated by a priori knowledge of distortions. Nevertheless, such methods rapidly become unreliable if a new type of distortion is encountered.

A more universal approach is to use a priori assumptions on the expected clean signal rather than on the distortions that may occur. In such scheme, the distortions are characterized by comparing some properties of the input signal with an a priori model of these properties for clean signal. This scheme allows to deal with a wide range of distortion type, we call it the "source-based approach". For example, Gray [6] uses a speech production model to find parts of the incoming acoustic signal that cannot be produced by the human vocal-tract. In a same way, Jin [5] learns a codebook of clean speech spectral parameters and uses it as a reference for computing objective distances measures.

However, in order to predict the subjective impact of degradations, one should characterize them using perceptually relevant features. Therefore, we propose to apply the source-based approach in the perceptual domain. We model a set of perceptual features extracted after an auditory transform of clean speech. In this way, any distortion of a processed signal is characterized by the deviation from expected value for each perceptual feature. These perceptual features deviations are then mapped to subjective opinion. Such method may be compared to the feature calculation used in [7] to interpret audible errors. The detailed scheme of the proposed non-intrusive assessment technique is presented in the next section and preliminary results are given to illustrate the characterization of distortions by some perceptual features.

3. Perceptually motivated approach

A schematic diagram of our measurement system is shown Figure 1. The acoustic signal is filtered by a peripheral ear model that simulates key auditory properties like frequency resolution, masking and compression. The output of the peripheral ear model is a time-frequency representation, which is expected to carry only audible aspects of the signal. This representation is further processed to extract a set of perceptual features like timbre, periodicities, amplitude modulations. These features are made

explicit by separate time-frequency representations. A second stage compares observed perceptual features with expected features given a source model. The result of this comparison is the observed deviation from expected value for each perceptual feature. At the last stage, these deviations are mapped to the subjective opinion, using a learned non-linear relationship.

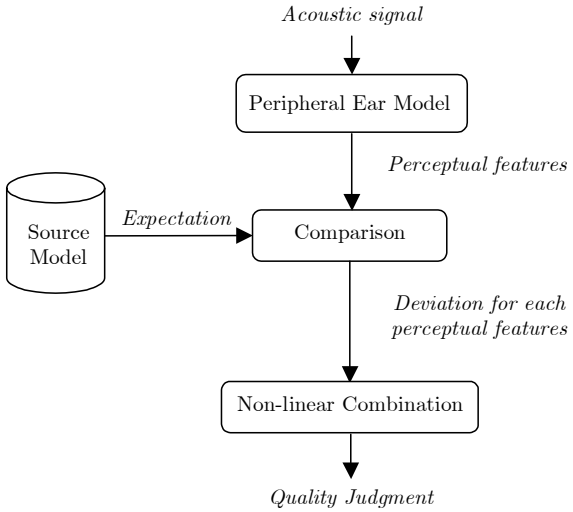

Figure 1: Source-based scheme of quality assessment.

One may draw interesting parallel between our source-based approach and some models of perceptual streaming. According to Ellis [8], the auditory system uses an internal "world model" to organize auditory input into separate objects and to identify them. The world model may represent a priori knowledge at different level of abstraction, from perceptual features of speech or other familiar acoustic events to linguistic knowledge or context knowledge. Our approach assumes that a perceived distortion corresponds to an unknown or unexpected object. As we cannot model linguistic knowledge, the non-intrusive measure should be compared to subjective opinion of listeners assessing a degraded speech of foreign language without hearing the original.

The development of a source model is an on going work. We discuss below the calculation of some perceptual features.

4. Calculation of perceptual features

4.1. Auditory periphery model

The auditory model consists of the following stages:

- outer and middle ear filters
- 20 band ERB-spaced Gammatone filter
- amplitude rectification
- adaptive compression

The Gammatone filterbank have been optimized to match the Equivalent Rectangular Bandwidth (ERB) of the auditory system [9]. The use of a filterbank to model frequency resolution gives advantages in term of temporal resolution. Although intrusive measurements like [1] consider only the energy spectrum of speech, many psychoacoustical experiments have demonstrated the importance of time in perception [10]. We rectify the filters output by computing their instantaneous amplitude in a similar way as in [7].

The adaptive compression is the second major component of peripheral ear model. It accounts for effects of temporal adaptation and dynamic compression. This function is performed by four stages of Automatic Gain Control (AGC), each with a different time constant [10]. Cross-channel coupling between AGC simulates frequency suppression and the ear's adaptation to spectral tilt.

The output of the peripheral ear model is a time-frequency representation illustrated Figure 2.c. It represents compressed instantaneous amplitude versus time and ERB-rate scale (a scale similar to the Bark scale). It is compared to a spectrogram representation (log-energy versus time and linear frequency).

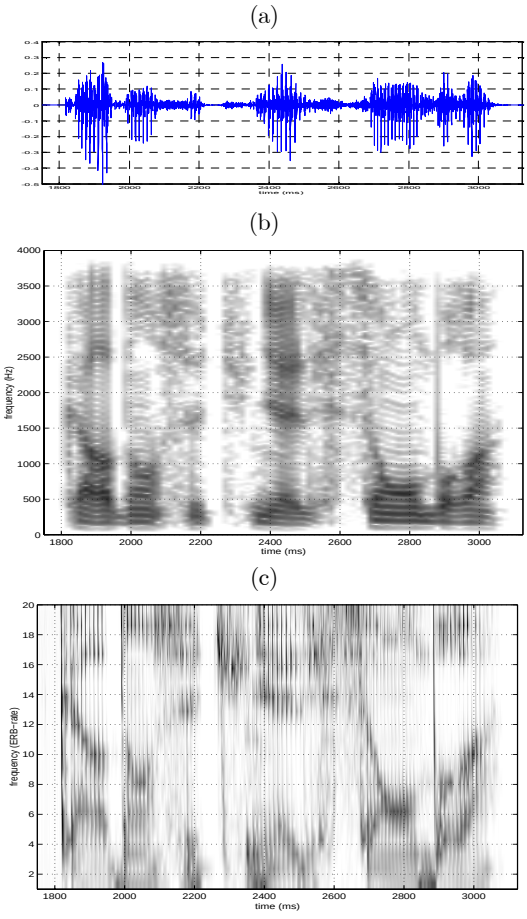

Figure 2: Peripheral ear representation (c) vs. spectrogram representation (b) of signal (a).

As an effect of the ear selectivity, harmonics have been smoothed in the peripheral ear representation. However, the fine time structure of the filterbank outputs exhibits signal transients like plosives as well as long-term periodicities. Therefore, one must exploit this temporal structure in order to extract all perceptually relevant information. The effects of adaptive compression are also clearly apparent, since signal onsets are enhanced in the peripheral ear representation.

4.2. Perceptual features

We have derived three perceptual features from peripheral ear representation $A_k(t)$: timbre, pitch and envelope modulation. The timbre representation is simply determined by low pass filtering the peripheral ear representation $A_k(t)$ with a cut-off frequency w_c of 25Hz. The periodicity of the high-passed $A_k(t)$ with same cut-off frequency w_c gives the pitch. Envelope modulations characterize temporal structure of the timbre representation $A_k^L(t)$. It is defined by:

$$mod(f,t) = \frac{\left| dA_k^L(t)/dt \right|}{1 + \alpha A_k^L(t)} \qquad (1)$$

We do not claim that these features are able to characterize all occurring distortions. This is clearly a non-exhaustive set of features. We simply want to illustrate the usefulness of these features for some non-linear distortions introduced by low bit-rate coders. This is the aim of the next section.

5. Characterizing distortions

We consider here GSM distorted speech signals and study how theirs distortions may be characterized by using a priori assumptions on perceptual features.

Let us consider the distorted signal illustrated Figure 3. It shows clipped parts occurring nearly between 200-300ms and between 550-650ms. There is also a muted part between 800-1000ms. Listeners can easily detect clipped parts using their linguistic knowledge but also because the beginning of speech after a clipping seems "abrupt". This latter distortion can be characterized by the envelope modulation.

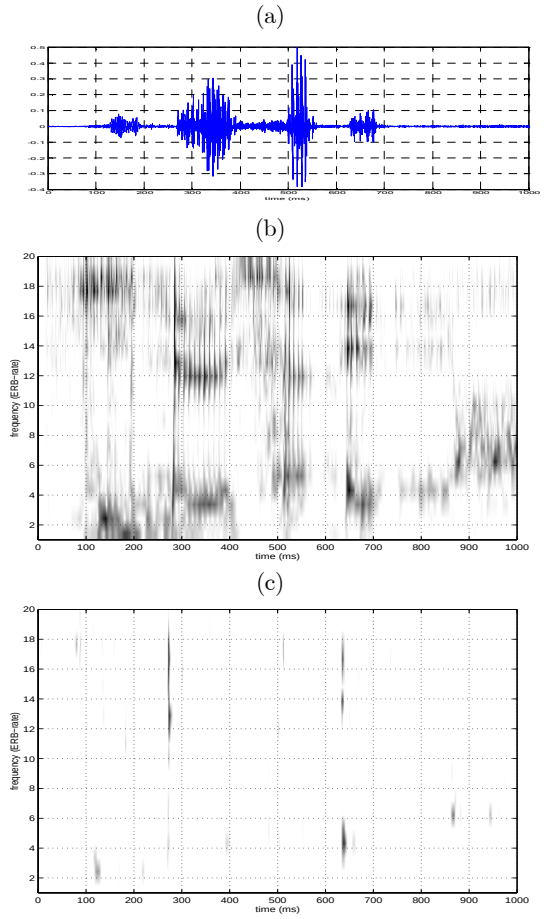

Figure 3: Distorted signal (a); Peripheral ear representation (b) and Envelope modulation (c).

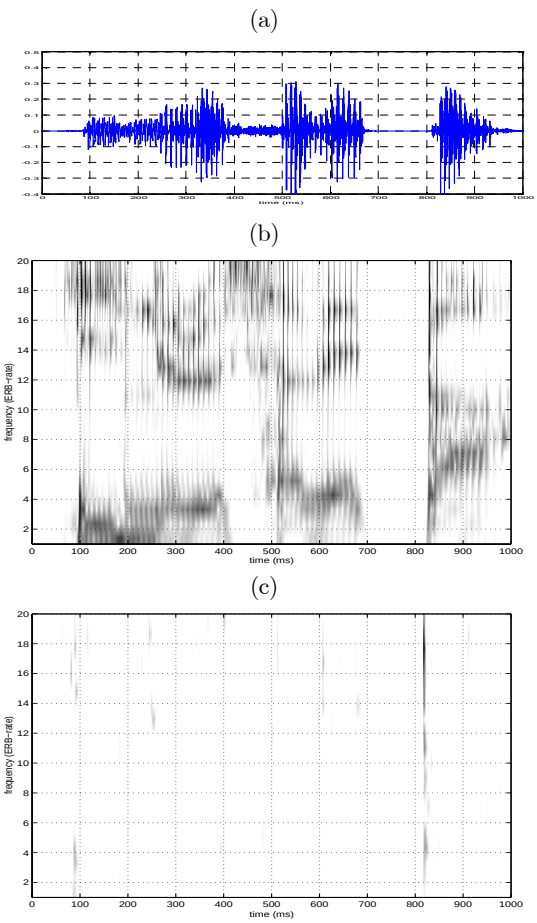

Figure 4: Original signal (a); Peripheral ear representation (b) and Envelope modulation (c).

We can see on Figure 3.c that high values of the envelope modulation $mod(f,t)$ coincide with the beginning of speech after a clipping and with the peaks of $A_k^L(t)$ at that time. However, envelope modulation of the clean speech shows equivalent high values near the time instant t=800ms, which corresponds to a plosive. In order to make a reliable detection, we must consider the value and the distribution of $mod(f,t)$ along the ERB scale. In this way, we could classify high values of modulation as corresponding to "plosives" (Figure 4.c) or to "vowels" (Figure 3.c). We assume that a vowel showing high envelope modulation will be perceived as a distortion.

Another example of distorted speech is shown Figure 5. A degradation occurs between 4050-4100ms and is subjectively assessed as "metallic voice". Let us consider the fine structure of the peripheral ear representation $A_k(t)$. Between 4050-4100ms, high frequency components exhibit temporal periodicities whereas low-frequency components have no periodicities or harmonic structure. The pitch salience confirms that long-term periodicities are audible. Therefore, under the realistic assumption that low frequency harmonics dominate in voiced clean speech, we are able to characterize such distortion.

The examples showed here do not assert the reliability of our assessment scheme. We presented them to illustrate what we consider a promising approach.

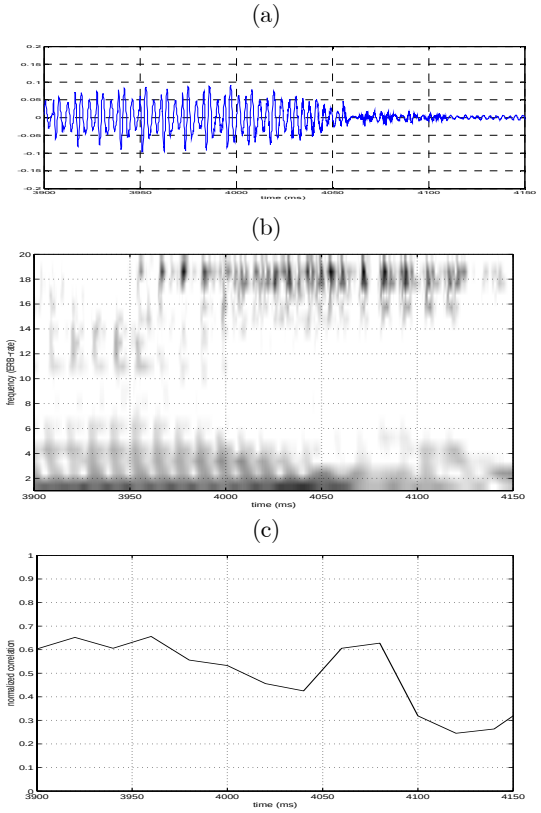

Figure 5: Distorted signal (a); Peripheral ear representation (b) and Pitch salience (c).

6. Conclusion

The major issue of non-intrusive assessment is to characterize distortions. We proposed here a source-based approach in the perceptual domain. The temporal resolution of our perceptual representations seems to provide interesting cues to characterize non-linear distortions induced by low bit-rate coders. However, a larger set of perceptual features has to be found and modeled.

7. References

[1] Perceptual Evaluation of Speech Quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, ITU-T Draft Recommendation P.862, May 2000.

[2] Calculation of loudness ratings for telephone sets, ITU-T Recommendation P.79, 1993.

[3] BT: Call Clarity Index verification, ITU-T Study Group 12 question 15, delayed contribution, 1998.

[4] Veaux, C., Scalart, P. and Gilloire, A., Analysis and On-line Detection of Audible Distortions in GSM Telephony, in Eurospeech'99, vol. 6, pp. 2579-2582, 1999.

[5] Jin and Kubichek, Vector Quantization Techniques for Output-Based Objective Speech Quality Measure, in Proc. ICASSP, 1996.

[6] Gray, P., Hollier, M.P., Massara, R.E., Non-Intrusive Speech Quality Assessment Using Vocal-Tract Models, IEE Proc. on Vision, Image and Signal Processing, vol. 147, Issue 6, Dec. 2000, pp. 493-501.

[7] Thiede, T., Treurniet, W.C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J.G., Colomes, C., Keyhl,M., Stoll, G., Brandenburg, K. and Feiten, B. "PEAQ – The ITU standard for objective measurment of perceived audio quality". *Journal of the A.E.S.*, 48 (1/2), 3-29, Jan/Feb 2000.

[8] Ellis, D.P.W. "Using knowledge to organize sound: the prediction-driven approach to computational auditory scene analysis and its application to speech/non-speech mixtures" *Speech Communication*, 27, 3-4,,pp.281-298, April 1999.

[9] Patterson, R., Nimmo-Smith, I., Holdsworth, J. and Rice, P. "An efficient auditory filterbank based on the gammatone function" *Appendix B of SVOS Final report: The Auditory Filterbank*. APU report 2341,1987.

[10] Slaney, M., Lyon, R.F. "On the importance of time – a temporal representation of sound", *Visual Representations of Speech Signals*, Cooke, M. and Crawford, M. (eds.), Wiley, pp. 95-116, 1993