



FEATURES INTERPOLATION DOMAIN FOR DISTRIBUTED SPEECH RECOGNITION AND PERFORMANCE FOR ITU-T G.723.1 CODEC

Vladimir Fabregas Surigué de Alencar and Abraham Alcaim

Center for Telecommunications Studies (CETUC)
Pontifical Catholic University of Rio de Janeiro (PUC-RIO), Rio de Janeiro, Brazil
 vladimir@cetuc.puc-rio.br, alcaim@cetuc.puc-rio.br

ABSTRACT

In this paper, we examine the best domain to perform features interpolation in Distributed Speech Recognition (DSR) systems. We show that the only one domain where a performance gain can be achieved from the linear interpolation procedure is in the Line Spectral Frequencies (LSF) domain. A DSR scenario where the ITU-T G.723.1 codec is employed is also investigated. The recognition feature generated from the reconstructed speech is highly sensitive to the encoding noise. We have also shown that the LSF quantization scheme used by the G.723.1 codec decreases the recognition performance by approximately 2 %.

Index Terms— Speech recognition, ITU, Linear predictive coding, Interpolation, Hidden Markov models

1. INTRODUCTION

The growth of the Internet and cellular mobile communication networks, along with the increasing interest in Automatic Speech Recognition (ASR) systems, have stimulated the development of Distributed Speech Recognition (DSR) services. Such services perform ASR in a server system, based on the acoustic parameters extracted at the user terminal. This procedure allows that the high complexity and large memory requirements of ASR systems be distributed between the simple/low power client devices and the remote server.

Most speech coders employed in mobile communication systems and IP networks operate at low bit rates and utilize, in general, LPC (Linear Predictive Coding) algorithms based on a speech production model. In this model, an excitation signal is applied to an all-pole filter (characterized by the LPC parameters), that represents the spectral envelope information of the speech signal. Usually, the LPC parameters are transformed to LSF (Line Spectral Frequencies), due to attractive properties of the latter to the quantization and interpolation procedures. It is also known that in DSR systems, extracting speech recognition features from the parameters of a speech coder provides better recognition performance than obtaining the features from the decoded/reconstructed signal [1]. However, the parameters of a speech coder are not the most adequate ones for the remote recognition system. For this reason, different codec parameter transformations have to be considered in order to improve recognition accuracy.

Another important remark is that for satisfactory operation of the ASR system, the recognition features have to be obtained at a high rate (typically 100 Hz). However, speech coders for mobile

telephony and IP networks generate their parameters at lower rates (e.g., 50 Hz or 33 Hz). In a recent study on the efficiency of recognition features for distributed speech recognition [2], it was shown that low rates significantly degrade the performance of the recognizer. Hence, it is paramount to interpolate the speech features in order to achieve a recognition performance which is closer to the one obtained when the features are extracted at a high rate. Another major point is concerned to the domain where the interpolation will be carried out. Is it a good procedure to interpolate in the domain of the recognition features? Or in the domain of the LPC parameters? Or, is it better to interpolate in the domain of the LSF parameters? The answer to these questions is not available in the literature and is the main purpose of this paper. Results obtained with the ITU-T G.723.1 codec [3] in a DSR scenario are also presented and discussed.

2. RECOGNITION FEATURES

The recognition features can be extracted directly from the LPC parameters, without the need to reconstruct the speech signal. In speech decoders, these parameters are obtained in a stage before speech reconstruction. This means that recognition features extracted in this stage are less complex than the ones obtained from the reconstructed speech, since they avoid the need of speech recovery. Moreover, it is important to remark that generating features from the reconstructed speech at the decoder yields worse recognition performance than directly extracting them from the codec parameters. Recognition features that can be obtained from the LPC parameters are the LPCC (LPC Cepstrum) and MLPCC (Mel-Frequency LPCC) [4].

The Line Spectral Frequencies (LSFs) are often used in speech coders due to their high coding efficiency and their attractive interpolation properties [5]. Extracting recognition features from the LSFs avoids a speech decoding operation, as well as a conversion of LSF to LPC. A distributed speech recognition system that adopts this strategy becomes computationally more efficient than any other one based either on speech reconstruction or on LPC parameter transformations. The recognition features which can be obtained from the LSFs are the PCC (Pseudo-Cepstral Coefficients) [6], MPCC (Mel-Frequency PCC) [6], PCEP (Pseudo-Cepstrum) [1] and MPCEP (Mel-Frequency PCEP) [1]. It is worth to mention that these features, which are directly obtained from LSFs, correspond to approximations of the LPCC and MLPCC features obtained from LPC parameters. Note that the use of these approximations avoid the need to recover LPC parameters to obtain the recognition features.

In this paper, we will consider only the MEL scale features (MLPCC, MPCC and MPCEP), since they provide a much better performance than the ones achieved with the linear scale features (LPCC, PCC and PCEP) [2]. The MFCC (Mel-Frequency Cepstral Coefficients) [7]-[8] features will be also obtained from voice reconstructed with the G.723.1 codec [3].

2.1. MEL-FREQUENCY LPCC (MLPCC)

The extraction process of the LPCC features from the LPC coefficients is formulated in the z-transform domain, using the complex logarithm of the LPC system transfer function, which is analogous to the cepstrum computation from the discrete Fourier transform of the speech signal [9]. The i -th LPCC parameter is given by the following recursive equation

$$c_i = \begin{cases} \ln(G) & i = 0 \\ a_1 & i = 1 \\ a_i + \sum_{j=1}^{i-1} \frac{i-j}{i} c_{i-j} a_j & 1 < i \leq p \\ \sum_{j=1}^p \frac{i-j}{i} c_{i-j} a_j & i > p \end{cases} \quad (1)$$

where a_i is the i -th LPC parameter, p is the LPC system order and G is the gain factor of the system.

The MLPCC feature is obtained by transforming the real frequency axis of the LPCC to the mel frequency scale. This is performed by a bank of n first-order all-pass filters, where n is the number of LPCC features [10]. The filters have their first-order all-pass transfer function $\psi(z)$ [9] given by

$$\psi(z) = \frac{z^{-1} - a^*}{1 - az^{-1}} \quad (2)$$

where a is the all-pass filter coefficient and a^* is the complex conjugate of a . Each LPCC parameter, c_i , is processed by a different filter.

Since the purpose of each filtering operation is to approximate the mel scale frequency, it is important to analyze the relationship of the transfer function given by (2) and the transformation of the frequency axis. In order to simplify the filter implementation, let a be a real number [11]. Now rewrite ψ , as a function of $e^{j\Omega}$, as

$$\psi(e^{j\Omega}) = e^{-j\theta(\Omega)} \quad (3)$$

where Ω is the real frequency. From (2) and (3), we can derive the mel scale frequency as a function of the real frequency Ω :

$$\theta(\Omega) = \arctan \left[\frac{(1-a^2)\sin\Omega}{(1+a^2)\cos\Omega - 2a} \right] \quad (4)$$

Changing the value of a it is possible to adjust $\theta(\Omega)$ to the mel scale curve. At an 8kHz sampling frequency, the value of a that best approximates the mel scale curve is 0.3624 [11].

The outputs of the filter bank are the MLPCC features.

2.2. MEL-FREQUENCY PCC (MPCC)

To obtain the MPCC features from the PCC [6], the LSFs w_i are replaced by w_i^m , which are defined by the transformation

$$w_i^m = w_i + 2 \tan^{-1} \left(\frac{0.45 \sin w_i}{1 - 0.45 \cos w_i} \right) \quad (5)$$

This expression transforms the frequency axis of a particular set of parameters to the mel scale frequency axis [12]. The MPCC features are expressed by

$$\hat{c}_n^m = \frac{1}{2n} (1 + (-1)^n) + \frac{1}{n} \sum_{i=1}^p \cos n w_i^m \quad (6)$$

where \hat{c}_n^m is the n -th MPCC.

2.3. MEL-FREQUENCY PCEP (MPCEP)

Following the same procedure described for the MPCC, we can express the MPCEP features by

$$\hat{d}_n^m = \frac{1}{n} \sum_{i=1}^p \cos n w_i^m \quad (7)$$

where \hat{d}_n^m is the n -th MPCEP.

3. FEATURES INTERPOLATION RESULTS

In the simulations carried out in this section, the speech frames have 25 ms duration and the frame rate is either 100 Hz or 50 Hz, depending on the desired rate of the LPC or LSF extractions. The 100 Hz frame rate was chosen because this is the usual value employed by speech recognizers to provide good performance. The 50 Hz frame rate was chosen because this value is usual in voice coders operating in IP networks and mobile environments.

It is important to note that the only features that will be vector quantized by the codec schemes in a distributed system are the LSF parameters. The quantized (and received) LSF parameters in the remote terminal will then be converted to the appropriate speech recognition features. In this section, however, we have used unquantized LSF parameters. We have found that in a first set of experiments it would be interesting and more advisable to know the particular behavior of the interpolation domain of the features obtained from transformations of the original LPC and LSF parameters to speech recognition features. This will certainly bring some benefits (and eventually a better insight) during the analysis of these features when operating in the distributed systems and using the several codec processing schemes of mobile and IP environments. In the next section, we will consider the situation where the G.723.1 codec is used in the distributed recognition system.

In all experiments in this work, the feature extractors generate one set of 10 parameters plus its derivatives (parameters) representing a total of 20 recognition features. The ASR system

considered in our experiments is a speaker-independent, isolated word recognizer. The speech database is composed of 50 male speakers and 50 female speakers, each one repeating three times the digits 0,1,2,3,4,5,6,7,8,9 and the word “meia” in Portuguese. This represents a total of 3,300 words. A distribution of 70% and 30% of the speech database was used for training and testing, respectively. Training and testing were always used on matched conditions. The recognition systems use five-state continuous observation HMMs (Hidden Markov Models) with a mixture of three Gaussians per state. They were implemented with the HTK (HMM Toolkit) software [7].

Table 1 shows the recognition results achieved with for the linear interpolation of the recognition features (from 50 Hz to 100 Hz) in their own domain, in the LPC domain and in the LSF domain. It is also shown the recognition accuracy obtained without interpolation at the rates of 100 Hz and 50 Hz. It should be remarked that in each test case, the model parameters are trained with the same type of features (same type of interpolation), i.e., training and testing are matched in this sense.

TABLE 1: RECOGNITION ACCURACY IN DIFFERENT INTERPOLATION DOMAINS

Interpolation Domain - Rate	MLPCC	MPCC	MPCEP
No Interpolation - 100 Hz	98.3%	97.5%	98.2%
No Interpolation - 50 Hz	93.8%	93.1%	93.7%
Feature Domain - 100 Hz	93.9%	93.8%	94.4%
LPC Domain - 100 Hz	93.8%	no	no
LSF Domain - 100 Hz	96.0%	95.7%	96.0%

In order to interpolate the recognition features (MLPCC, MPCC and MPCEP) in their own domains, we first generate them from the LSF and LPC parameters at 50 Hz and then interpolate to the 100 Hz rate.

Comparing the performance of the interpolation in the features domain at the 100 Hz rate (3rd row of Tab. 1) with the ones obtained without interpolation at the rate of 50 Hz (2nd row of Tab. 1), it can be seen that the use of interpolation in the features domain does not bring any significant gain. It can also be verified that the interpolation in the LPC domain does not offer any performance gain for the unique parameter that can be interpolated in this domain.

It is now interesting to compare the performance of the features obtained at the 100 Hz rate from the linear interpolation of the LSF parameters (last row of Tab. 1) and the features obtained at the 50 Hz rate without interpolation (2nd row of Tab. 1). It can be seen that the improvements of the recognition rates are 2.2%, 2.6% and 2.3% for the MLPCC, MPCC and MPCEP, respectively, when the interpolation are carried out in the LSF domain. Consequently, this is the interpolation domain for which the results are the nearest ones to those obtained when the features are generated at 100 Hz (ideal situation – 1st row of Tab.1). For distributed speech recognition, this means that the best option is to first interpolate at the LSFs and then generate the features that will be used in the recognition system.

4. ANALYSIS USING THE ITU-T G.723.1 CODEC

In this section, we consider the interpolation problem in a distributed speech recognition system using the ITU-T G.723.1 codec. Figure 1 illustrates this DSR scenario. The ITU-T G.723.1

codec is one of the most widely used standards for IP networks nowadays. It allows speech encoding at 6.3 Kbit/s or 5.3 Kbit/s. In our experiments we have used the 6.3 Kbit/s operation mode. The G.723.1 codec employs 30 ms-frames, 8 KHz sampling rate, and 10 LSFs per frame. The LSFs are quantized with a 24 bit predictive split vector quantizer and transmitted at a 33Hz rate (one every 30 ms). Therefore, interpolating the LSFs from 1 per 30 ms to 1 per 10 ms is equivalent to an interpolation by a factor of 3. Based on the results presented in the previous section, we have only considered interpolation in the LSF domain. This means that the features based on LSF (MPCC and MPCEP) or LPC (MLPCC) will be obtained at 100 Hz by the linear interpolation of the LSF parameters from 33Hz to 100 HZ. The MFCC feature is generated from the reconstructed speech. Hence, no interpolation will be required, since this feature can be directly extracted from the decoded speech at the 100 Hz rate.

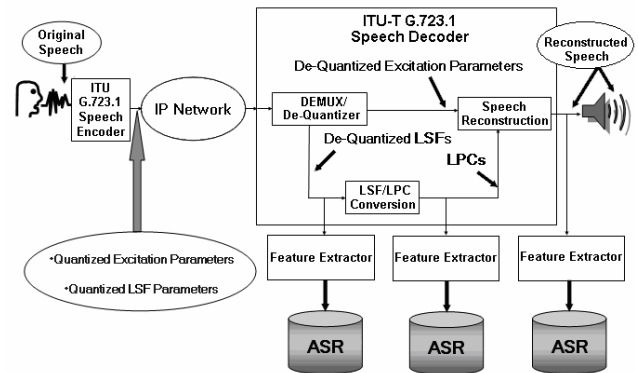


Figure 1: Codec Features extractors and ASR systems using the ITU-T G.723.1

We have examined the impact of interpolating the unquantized LSFs and the LSFs quantized by the G.723.1 codec on the recognition accuracy obtained at the remote ASR system. It should again be remarked that in each test case, the model parameters are trained with the same type of features (quantized vs quantized, etc). The results are presented in Table 2. We have also obtained the recognition performance of the MFCC feature extracted from the reconstructed speech at the 100 Hz rate. As compared to the results presented in Table 2, the performance of this feature – 88.1% – is even worse than the ones obtained with any of the features without interpolation. This shows that the MFCC is very sensitive to the encoding noise, since the performance obtained with this parameter from the original speech is 99.4 %.

TABLE 2. RECOGNITION ACCURACY USING THE ITU-T G.723.1 CODEC

Interpolation Domain - Rate	MLPCC	MPCC	MPCEP	MFCC
No Interp. - Quantized LSF - 33 Hz	89.4%	88.7%	89.2%	88.1%
Quantized LSF - 100 Hz	91.9%	91.8%	92.3%	no
No Interp. - Unquantized LSF - 33 Hz	91.5%	89.9%	91.3%	no
Unquantized LSF - 100 Hz	94.1%	93.8%	94.2%	no

In the DSR scenario using the ITU-T G.723.1 codec, the use of interpolation provides a performance gain of 3% on the average. It can also be seen from Table 2 that the MPCEP feature outperforms the other ones, besides requiring a lower computational complexity. Moreover, comparing the 2nd and 4th rows of Table 2, we verify that the LSF quantization noise due to

the G.723.1 codec causes a significant decrease of the recognition performance (considering the 100 Hz rate): 2.2 % for the MLPCC, 2.0 % for MPCC and 1.9 % for MPCEP.

5. CONCLUSION

In this paper, we have shown that the best domain to perform linear interpolation of the recognition features used in distributed speech recognition is the LSF domain. The results are significantly higher than the ones obtained from both the interpolation in the LPC domain and in the features domain. We have also investigated the distributed speech recognition system when the ITU-T G.723.1 codec is used. We have shown that the MFCC feature, which is obtained from the reconstructed speech, is highly sensitive to the encoding noise. The performance drops 11.3%. The features obtained from the LSF parameters can provide much better recognition accuracy, especially if they are interpolated to 100 Hz in the LSF domain. However, the LSF quantization scheme employed by the G.723.1 codec deteriorates the recognition performance by approximately 2 %.

6. REFERENCES

- [1] H. S. Choi, H. K. Kim, and H. S. Lee, "Speech Recognition Using Quantized LSP Parameters and their Transformations in Digital Communication", vol. 30, pp. 223-233, Speech Communication, 2000.
- [2] V. F. S. Alencar and A. Alcaim, "Transformations of LPC and LSF Parameters to Speech Recognition Features", Proceedings of the ICAPR, Bath, UK, August 2005.
- [3] ITU-T Recommendation G.723.1, "Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 Kbit/s," March 1996.
- [4] Y. Ohshima, "Environmental Robustness in Speech Recognition using Physiologically-Motivated Signal Processing," PH. D. Thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, December 1993.
- [5] W. B. Kleijn and K. K. Paliwal, Speech Coding and Synthesis, Amsterdam, The Netherlands: Elsevier, 1995.
- [6] H. K. Kim, S. H. Choi and H. S., Lee, "On Approximating Line Spectral Frequencies to LPC Cepstral Coefficients," IEEE Trans. Speech and Audio Processing, vol. 8, pp. 195 – 199, March 2000.
- [7] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, The HTK Book (for HTK Version 3.2.1), December 2002.
- [8] S. B. Davies and P. Mermelstein, "Comparasion of Parametric Representations for Mono syllabic Word Recognition in Continuously Spoken Sentences," vol.28, pp.357-366, IEEE Trans. ASSP, August 1980.
- [9] S. K. Mitra, Digital Signal Processing: A Computer-Based Approach, McGraw-Hill International Editions, (1998)
- [10] A. V. Oppenheim e D. H., Johnson, "Discrete Representation of Signals," Proc. IEEE, vol. 60, pp.681- 691, June (1972)
- [11] M. Wölfel, J. McDonough, e A., Waibel, "Minimum Variance Distortionless Response on a Warped Frequency Scale," Eurospeech, Geneva, 2003.
- [12] F. S. Gurgem, S. Sagayama, e S. Furui, "Line Spectrum Frequency-Based Distance Measures for Speech Recognition," pp.521-524, Proc. ICSLP, Kobe, Japan, November (1990)