Referências Bibliográficas

[1] Choi, H. S., Kim, H. K. and Lee, H. S., “Speech Recognition Using Quantized LSP Parameters and their Transformations in Digital Communication,” Speech Communication, vol. 30, pp. 223-233, 2000.

[2] Cirigliano, R. J. R., Monteiro, C., Barbosa, F. L., Resende Junior, F. G. V., Couto, L. R., Moraes, J. A., “Um Conjunto de 1000 Frases Foneticamente Balanceadas para o Português Brasileiro obtido utilizando a abordagem de Algoritmos Genéticos”, pp. 544-549 XXII Simpósio Brasileiro de Telecomunicações, 2005.

[3] Peinado, A. M., Sánchez, V., Pérez-Córdoba, J. L. and Torre, A., “HMM – based channel error mitigation and its application to distributed speech recognition,” Speech Communication, vol.41, pp. 549-561, March, 2003.

[4] Kim, H. K., Choi, S. H. and Lee, H. S., “On Approximating Line Spectral Frequencies to LPC Cepstral Coefficients,” IEEE Trans. Speech and Audio Processing, vol. 8, pp. 195 – 199, March 2000.

[5] Ohshima, Y., “Environmental Robustness in Speech Recognition using Physiologically-Motivated Signal Processing,” PH. D. Thesis, Carnegie Mellon University, Pittsburgh, Pennsylvanya, December 1993.

[6] Kim, H. K., and Cox, R. V., “A Bitstream-Based Front-End for Wireless Speech recognition on IS-136 Communications System,” IEEE Trans. Speech and Audio Processing, vol. 9, pp. 558-568, July, 2001.

[7] Davis, S. B. and Mermelstein. P., “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans. Acoust., Speech, Signal Processing, vol. 28, pp. 357-366, 1980.

[8] Shannon, B. J., Paliwal, K. K., “Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition,” Speech Communication, vol. 48, pp. 1458-1485, 2006.

[9] You, K.-H. e Wang, H.-C., "Robust Features Derived from Temporal Trajectory Filtering for Speech Recognition under the Corruption of Additive and Convolutional Noises," ICASSP'98, pp.577-580, 1998.

[10] Kim, D.-S., Lee, S.-Y. and Kil, R. M., “Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments,” IEEE Trans. Speech and Audio Processing, vol. 7, pp. 55- 69, January, 1999.

[11] Ghulam, M., et al, “Pitch-synchronous ZCPA (PS-ZCPA)-based feature extraction with auditory masking,” Proc. ICASSP05, vol. 1, pp. 517-520, 2005.

[12] Ghulam, M., Horikama, J., and Nitta, T. “A Pitch-synchronous peak-amplitude based feature extraction method for noise robust ASR,” Proc. ICASSP06, vol. 1, pp. 505-608, 2006

[13] Gajic, B. and Paliwal, K. K., "Robust feature extraction using subband spectral centroid histograms," IEEE Int. Conf. on Acoust., Speech and Signal Processing, vol. 1, pp. 85-88, Salt Lake City, USA, May, 2001.

[14] Baker, J. K., "The Dragon System – an overview," IEEE Trans. ASSP, vol. 23(1), pp. 24-29, 1975.

[15] Jelinek, F., "Continuous Speech Recognition by Statistical Methods," Proc. IEEE, vol. 64 (4), pp. 532-556, 1976.

[16] Tevah, R. T., "Implementação de um Sistema de Reconhecimento de Fala Contínua com Amplo Vocabulário para o Português Brasileiro," Dissertação de Mestrado, Junho de 2006.

[17] Cole, R. A., Mariani, J., Uszkoreit, H., Zaenen, A. and Zue, V., "Survey of the State of the Art in Human Language Technology," Cambridge University Press, Cambridge, UK, 1997, (http://cslu.cse.ogi.edu/HLTsurvey).

[18] Huang, X., Acero, A., Hon, H.-W., " Spoken Language Processing, A guide to Theory, Algorithm and System Development," Prentice-Hall, 2001.

[19] Baum, L. E., "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," Inequalities, vol. 1, pp. 1-8, 1972.

[20] Hwang, M. Y. and Huang, X., "Shared Distribution Hidden Markov Models for Speech Recognition," IEEE Trans Speech and Audio Processing, vol. 1(4), pp414-420, 1993.

[21] Young, S. J. and Woodland, P. C., "State Clustering in HMM-based Continuous Speech Recognition," Computer Speech and Language, vol. 8(4), pp. 369-384, 1994.

[22] Bahl, L. R., Souza, P. V., Gopalakrishnan, P. S., Nahamoo, D. and Picheny, M. A., "Context Dependent Modeling of Phones in Continuous Speech Using Decision Trees," Proc DARPA Speech and Natural Language Processing Workshop, pp. 264-270, Pacific Grove, Calif., Feb, 1991.

[23] Kannan, A., Ostendorf, M. and Rohlicek, J. R., "Maximum Likelihood Clustering of Gaussians for Speech Recognition," IEEE Trans. on Speech and Audio Processing, Vol. 2(3), pp. 453-455, 1994.

[24] Young, S. J., Odell, J. J. and Woodland, P. C., "Tree-Based State Tying for High Accuracy Acoustic Modeling", Proc. Human Language Technology Workshop, pp. 307-312, Plainsboro NJ, Morgan Kaufman Publishers, March, 1994.

[25] "Corpus de Extractus de Textos Eletrônicos Nilc/ Folha de São Paulo (Ceten-Folha)", (http://acdc.linguateca.pt/cetenfolha/), 14 November 2005.

[26] Katz, S. M., "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", IEEE Trans ASSP, Vol. 35(3), pp. 400-401, 1987.

[27] Ney, H., Essen, U., Kneser, R., "On Structuring Probabilistic Dependences in Stochastic Language Modeling", Computer Speech and Language, Vol. 8(1), pp. 1-38, 1994.

[28] Jelinek, F., “Up from Trigrams: the Struggle for Improved Language Models”, Proc. Eurospeech, pp. 1037-1040, Genoa, 1991.

[29] Bahl, L. R., Brown, P. F., de Souza, P. V., Mercer, R. L., “A Tree-Based Statistical Language Model for Natural Language Speech Recognition”, IEEE Trans ASSP, Vol. 37(7), pp. 507-514, 1989.

[30] Waegner, N. P., Young, S. J., “A Trellis-based Language Model for Speech Recognition”, Proc ICSLP, pp. 245-248, Banff, Canada, October, 1992.

[31] Lau, R., Rosenfeld, R., Roukos, S., “Trigger-based Language Models: a Maximum Entropy Approach”, Proc ICASSP’93, Vol. 2, pp. 45-48, Minneapolis, 1993.

[32] Black, E., Jelinek, F., Lafferty, J., Margeman, D. M., Mercer, R., Roukos, S., “Towards History-based Grammars: Using Richer Models for Probabilistic Parsing”, Proc DARPA, Spoken Language Workshop, pp.31-34, February, 1992.

[33] Deligne, S., Bimbot, F., “Language Modeling by Variable Length Sequences: Theoretical Formulation and Evaluation of Multigrams”, Proc ICASSP, Vol. 1, pp. 169-172, Detroit, 1995.

[34] Forney JR., G. D., "The Viterbi Algorithm," Proc. IEEE, vol. 61, pp. 268-278, March 1973.

[35] Vintsyuk, T. K., “Speech Discrimination by Dynamic Programming”, Kibernetika (Cybernetics), vol. 4 (1), pp. 81-88, Jan.-Feb. 1968.

[36] Lowerre, B. T., “The HARPY Speech Recognition System”, PhD Thesis in Computer Science Department, Carnegie Mellon University, 1976.

[37] Ney, H. J. and Ortmanns, S., “Dynamic Programming Search for Continuous Speech Recognition”, IEEE Signal Processing Magazine, pp. 64-83, 1999.

[38] Schwartz, R., et al., "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Speech Signals," Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1985, Tampa, FLA pp. 1205-1208.

[39] ITU-T Recommendation G.723.1, “Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s,” (http://www.itu.int/en/pages/default.aspx), March, 1996.

[40] 3GPP TS 26.071 V6.0.0, “Mandatory speech CODEC speech processing functions, AMR speech CODEC - General description,” (http://www.3gpp.org/), December, 2004.

[41] 3GPP TS 26.171 V6.0.0, “Speech codec speech processing functions, Adaptive Multi-Rate - Wideband (AMR-WB) speech codec - General description,” (http://www.3gpp.org/), December, 2004.

[42] ITU-T Recommendation G.722.2, “Wideband coding of speech at around 16 k/bits using adaptive multi-rate wideband (AMR-WB),” (http://www.itu.int/en/pages/default.aspx), Jully, 2003.

[43] ITU-T Recommendation G.712, “Transmission performance characteristics of pulse code modulation channels,” (http://www.itu.int/en/pages/default.aspx), November, 2001.

[44] ITU-T Recommendation G.723.1, “Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s - Annex A: Silence compression scheme,” (http://www.itu.int/en/pages/default.aspx), November, 1996.

[45] ITU-T Recommendation G.723.1, “Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s - Annex B: Alternative specification based on floating point arithmetic,” (http://www.itu.int/en/pages/default.aspx), November, 1996.

[46] 3GPP TS 46.001 V8.0.0, “Full rate speech; Processing functions,” (http://www.3gpp.org/), December, 2008.

[47] 3GPP TS 46.002 V8.0.0, “Half rate speech; Half rate speech processing functions,” (http://www.3gpp.org/), December, 2008.

[48] 3GPP TS 46.051 V8.0.0, “Enhanced Full Rate (EFR) speech processing functions; General description,” (http://www.3gpp.org/), December, 2008.

[49] 3GPP TS 26.094 V6.1.0, “Mandatory speech CODEC speech processing functions, Adaptive Multi-Rate (AMR) speech codec - Voice Activity Detector (VAD),” (http://www.3gpp.org/), July, 2006.

[50] 3GPP TS 26.092 V6.0.0, “Mandatory speech CODEC speech processing functions, Adaptive Multi-Rate (AMR) speech codec - Comfort noise aspects,” (http://www.3gpp.org/), December, 2004.

[51] 3GPP TS 26.091 V6.0.0, “Mandatory speech CODEC speech processing functions, Adaptive Multi-Rate (AMR) speech codec - Error concealment of lost frames,” (http://www.3gpp.org/), December, 2004.

[52] Hammer F., Reichl P., Nordstrom T., Kubin G., "Corrupted Speech Data Considered Useful", Acta Acustica, Vol. 90(6), pp. 1052-1060, November/December, 2004.

[53] Kondoz, A. M., “Coding for Low Bit Rate Communication Systems,” Digital Speech, John Wiley & Sons, Ltd. Chichester, UK, 1999.

[54] 3GPP TS 26.090 V6.0.0, “Mandatory speech CODEC speech processing functions, Adaptive Multi-Rate (AMR) speech codec – Transcoding functions,” (http://www.3gpp.org/), December, 2004.

[55] 3GPP TS 26.104 V6.1.0, “ANSI C code for the floating-point Adaptive Multi Rate (AMR) speech codec,” (http://www.3gpp.org/), March, 2004.

[56] 3GPP TS 26.073 V6.0.0, “ANSI C code for the Adaptive Multi Rate (AMR) speech codec,” (http://www.3gpp.org/), December, 2004.

[57] 3GPP TS 26.194 V6.0.0, “Speech codec speech processing functions, Adaptive Multi-Rate - Wideband (AMR-WB) speech codec - Voice Activity Detector (VAD),” (http://www.3gpp.org/), December, 2004.

[58] 3GPP TS 26.192 V6.0.0, “Speech codec speech processing functions, Adaptive Multi-Rate - Wideband (AMR-WB) speech codec - Comfort noise aspects,” (http://www.3gpp.org/), December, 2004.

[59] 3GPP TS 26.191 V6.0.0, “Speech codec speech processing functions, Adaptive Multi-Rate - Wideband (AMR-WB) speech codec - Error

concealment of erroneous or lost frames," (http://www.3gpp.org/), December, 2004.

[60] Bistritz, Y. and Pellerm, S. "Immittance Spectral Pairs (ISP) for speech encoding," IEEE Int. Conf. Acoustics, Speech, Signal Processing, Vol. 2, pp. 9-12, 1993.

[61] 3GPP TS 26.190 V6.1.1, "Speech codec speech processing functions, Adaptive Multi-Rate - Wideband (AMR-WB) speech codec – Transcoding functions," (http://www.3gpp.org/), July, 2005.

[62] 3GPP TS 26.204 V6.0.0, "Speech codec speech processing functions, Adaptive Multi-Rate - Wideband (AMR-WB) speech codec - ANSI-C code," (http://www.3gpp.org/), December, 2004.

[63] 3GPP TS 26.173 V6.0.0, "ANSI-C code for the Adaptive Multi Rate-Wideband (AMR-WB) speech codec," (http://www.3gpp.org/), December, 2004.

[64] Kutwak, A. B. "Análise da codificação LPC para sinais de fala," Projeto Final de Curso, UFRJ, 1999.

[65] Deller, J. R., Proakis, J. G. And Hansen, J. H. "Discrete-time processing of speech signals," MacMillan, 1993.

[66] Rabiner, L. R., and Juang, B. H. "Fundamentals of speech recognition," Prentice Hall, 1993.

[67] Oppenheim, A. V., and Johnson, D. H. "Discrete Representation of signals," Proc. IEEE, vol.60, pp.681- 691, June, 1972.

[68] Mitra, S. K., Digital Signal Processing: A Computer-Based Approach, McGraw-Hill International Editions, 1998.

[69] Wölfel, M., McDonough, J., and Waibel, A. "Minimum Variance Distortionless Response on a Warped Frequency Scale," Eurospeech, pp. 1021–1024, Geneva, 2003.

[70] Kleijn, W. B., and Paliwal, K. K. "Speech Coding and Synthesis", pp. 774 Amsterdam, The Netherlands: Elsevier, 1995.

[71] Itakura, F. "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals," J. Acoustic Soc. America, Vol. 57, S35(A), 1975.

[72] Alencar, V. F. S. and Alcaim, A. "Transformations of LPC and LSF Parameters to Speech Recognition Features", Proceedings of the ICAPR, pp. 522-528, Bath, UK, August 2005.

[73] Gurgen, F. S., Sagayama, S., and Furui, S. "Line spectrum frequency-based distance measures for speech recognition," ICSLP, pp.521-524, Kobe, Japan, November, 1990.

[74] Stevens, S. S., and J. Volkman, "The relation of pitch of frequency: A revised scale", Am. J. Psychol., vol.53, pp.329-353, 1940.

[75] Alencar, V. F. S., and Alcaim, A. "Features Interpolation Domain for Distributed Speech Recognition and Performance for ITU-T G.723.1 CODEC", Proceedings of the ICSLP, pp.1142-1145, Antwerp, BE, August 2007.

[76] Oetken, G., Parks, T. W., and Schüssler, H. W. “New Results in Design of Digital Interpolators,” IEEE Trans. On Acoustics, Speech and Signal Processing, ASSP-23, pp. 301-309, June 1975.

[77] Young, S. et al., “The HTK Book (for HTK Version 3.4),” (http://htk.eng.cam.ac.uk/), December 2006.

[78] Young, S. “An Application Toolkit for HTK (ATK 1.6),” (http://htk.eng.cam.ac.uk/develop/atk.shtml), June 2007.

[79] Järvinen, K. “Standardisation of the Adaptive Multi-rate Codec,” European Signal Processing Conference (EUSIPCO), pp. 1313-1316, Tampere, Finland, 4–8 Setembro 2000.

[80] Wang, J., Gibson, J., “Parameter interpolation to enhance the frame erasure robustness of CELP coders in packet networks”, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 745-748, 2001.

[81] Bolot, J. C., “Characterizing end-to-end packet delay and loss in the Internet”, Proc. ACM SIGCOMM, pp. 289-298, September, 1993.

[82] Haykin, S., "Neural Networks: A Comprehensive Foundation," ed.2, Prentice Hall, 1999.

Apêndice

Neste apêndice iremos apresentar na seção A.1., os dados técnicos dos equipamentos usados para a gravação da base e na seção A.2., as publicações relacionadas a esta tese doutorado até o momento da sua submissão.

A.1.
Informações Técnicas da Gravação da Base Alcaim - Alencar

As características técnicas dos equipamentos usados para a gravação da base de vozes utilizada nesta tese.

- Microfone: SM 58-LC - Shure
- Filtro anti-puff: Shure
- Pré-amplificador: Mic200 Phantom Power
- Placa de Som: Sound Blaster X-Fi Xtreme Áudio
- Software para gravação e edição: Sony Sound Forge 8.0

A sala de gravação possui tratamento acústico, e os equipamentos de gravação, como computadores, pré-amplificadores, que podem emitir algum tipo de ruído, ficam fora da mesma. A sala permite a permanência apenas da pessoa que está realizando a gravação, minimizando assim também outras fontes de ruído.

Os Locutores desta base têm idades variando de 17 à 65 anos, sendo alguns deles profissionais na realização de locuções, outros treinados para a gravação de livros falados e outros sem nenhum conhecimento teórico ou prático sendo orientados durante a gravação.

As frases utilizadas pelos locutores são todas afirmativas, não existindo frases interrogativas, exclamativas, etc.

A.2. Publicações Relacionadas à Tese

- Alencar, V. F. S., e Alcaim, A., “LSF and LPC - Derived Features for Large Vocabulary Distributed Continuous Speech Recognition in Brazilian Portuguese,” Asimolar, pp. 1237-1241, 2008.
- Alencar, V. F. S., e Alcaim, A., “Digital filter interpolation of decoded LSFs for distributed continuous speech recognition,” Electronics Letters, Vol. 44, N. 17, pp. 1039-1040, August, 2008.
- Alencar, V. F. S., e Alcaim, A., “On the Performance of ITU-T G.723.1 and AMR-NB Codecs for Large Vocabulary Distributed Speech Recognition in Brazilian Portuguese” In: 2nd International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2009), Londres, 2009.
- Alencar, V. F. S., e Alcaim, A., “Recuperação de Pacotes Perdidos em Sistemas de Reconhecimento de Voz Distribuído Usando Redes Neurais,” In: XXVIII Simpósio Brasileiro de Telecomunicações, Blumenau - SC, 2009.

LSF and LPC - Derived Features for Large Vocabulary Distributed Continuous Speech Recognition in Brazilian Portuguese

V. F. S. Alencar and A. Alcaim
vladimir@cetuc.puc-rio.br, alcaim@cetuc.puc-rio.br
CETUC – PUC/RIO
22453-900, Rio de Janeiro/RJ, Brazil

Abstract - In this paper, we describe several important experiments concerning Large Vocabulary Distributed Continuous Speech Recognition (LVDCSR) systems in Brazilian Portuguese using LSF and LPC - Derived Features. The ITU-T G.723.1 codec is employed and investigated as a case of practical use of this technology. Results are presented for both speaker dependent and independent modes as well as the situations where the same text or different texts were used for training and testing.

I. Introduction

The growth of the Internet and cellular mobile communication networks, along with the increasing interest in more natural Automatic Speech Recognition (ASR) systems, have stimulated the development of Large Vocabulary Distributed Continuous Speech Recognition (LVDCSR) 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 LVDCSR 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. In this paper we have used linear interpolation on the LSF domain [3] to obtain features at 100 Hz from the ITU-T G.723.1 codec [4] in a LVDCSR scenario in order to improve the recognition performance of the system.

In Section II, we present a brief description of the recognition features used in the experiments reported in this paper. Section III describes the database and the experimental procedures. Simulation results are presented and discussed in section IV. Finally, Section V concludes de paper.

II. 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) [5].

The Line Spectral Frequencies (LSFs) are often used in speech coders due to their high coding efficiency and attractive interpolation properties [6]. 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) [7], MPCC (Mel-Frequency PCC) [7], 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) features [8]-[9] will be also obtained from voice reconstructed with the ITU-T G.723.1 codec at the two different operation rates (6.3 kbit/s or 5.3 kbit/s) [4]. The difference between the two rates of ITU-T G723.1 affects the MFCC because only the excitation encoding differ between the two rates. Therefore, this difference affects only the features derived from reconstructed speech – the MFCC.

A. 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 [10]. The *i*-th LPCC parameter is given by the following recursive equation

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

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 [11]. The filters have their first-order all-pass transfer function $\psi(z)$ [10] given by

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

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 [12]. Now rewrite ψ, as a function of $e^{j\Omega}$, as

$$\psi\left(e^{j\Omega}\right) = e^{-j\theta(\Omega)} \qquad (3)$$

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:

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

Changing the value of a it is possible to adjust $\theta(\Omega)$ to the mel scale curve. At an 8 kHz sampling frequency, the value of *a* that best approximates the mel scale curve is 0.3624 [12].

The outputs of the filter bank are the MLPCC features.

B. Mel-Frequency PCC (MPCC)

The PCC is computed directly from the LSFs. However, its derivation is based on the LPCC. Mathematical manipulations and approximations allow it to be expressed in terms of the LSFs [7]. The n-th PCC is given by the equation

$$\hat{c}_n = \frac{1}{2n}\left(1 + (-1)^n\right) + \frac{1}{n}\sum_{i=1}^{p} \cos nw_i \qquad (5)$$

where w_i is the i-th LSF parameter.

To obtain the MPCC features from the PCC [7], the LSFs w_i are replaced by w_i^m, wich are defined by the transformation

$$w_i^m = w_i + 2\tan^{-1}\left(\frac{0.45\, sin\, w_i}{1 - 0.45\cos w_i}\right) \qquad (6)$$

This expression transforms the frequency axis of a particular set of parameters to the mel scale frequency axis [13]. The MPCC features are expressed by

$$\hat{c}_n^m = \frac{1}{2n}\left(1 + (-1)^n\right) + \frac{1}{n}\sum_{i=1}^{p} \cos nw_i^m \qquad (7)$$

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

C. Mel-Frequency PCEP (MPCEP)

Using the mathematical expression of the PCC features, it is somewhat trivial to obtain the PCEP [1]. They are derived from the PCC by eliminating the $\frac{1}{2n}\left(1+(-1)^n\right)$ term. Note that this term does not depend on the speech signal, i.e., it does not depend on the LSF parameters. The n-th PCEP expression is given by

$$\hat{d}_n = \frac{1}{n}\sum_{i=1}^{p}\cos nw_i \qquad (8)$$

It is fair to expect a good spectral performance of the PCEP because they provide a spectral envelope very similar to the one provided by the Cepstrum, wich is generated from the original speech signal [1]. The PCEP features have the advantage of presenting a computational load even lower than the PCC.

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

$$\hat{d}_n^m = \frac{1}{n}\sum_{i=1}^{p}\cos nw_i^m \qquad (9)$$

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

III. Description of the Database and Experiments

The new database used in this paper was designed based on [14] that specifies a phonetically balanced set of 1000 sentences for the Brazilian Portuguese. The speech database is composed of 50 male speakers and 50 female speakers, each one repeating once all the 1000 sentences (3528 words). The database was recorded in a studio with 16 kHz sample rate and 16 bits per sample with a bandwith from 50 – 7000 Hz. This database was filtered and down sampled [10] to match the ITU-T G723.1 [4] requirements. Fig. 1 is a graphical representation of this database and will be used to explain the set of experiments carried out in this paper.

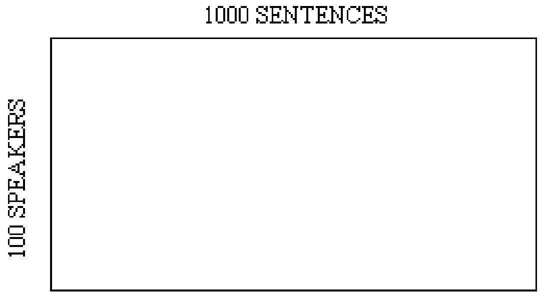


Fig. 1. Graphical representation of the constructed database

In the experiments, we consider LVDCSR systems using the ITU-T G.723.1 codec, as illustrated in Fig. 2. 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 considered both operation modes. 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 33 Hz rate (one every 30 ms). The 100 Hz frame rate was chosen because this is the usual value employed by speech recognizers to provide good performance. Therefore, interpolating the LSFs from 1 per 30 ms to 1 per 10 ms is equivalent to a linear interpolation by a factor of 3. Based on the results presented in [3], 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 original and reconstructed speech with 25 ms frame duration (with frames overlay so that parameters are generated at each 10 ms). Hence, no interpolation will be required, since this feature can be directly extracted from the original and decoded speech respectively at the 100 Hz rate.

It should be remarked that in all cases, the model parameters are trained with the same type of features that will be used in tests. This means that we are working in a matched condition.

The database was divided in three different ways to produce three different sets of experiments. The first set of experiments, represented by Fig. 3, can be considered a 100 speaker dependent scenario. The second set of experiments, represented by Fig. 4, can be considered a speaker independent scenario with all the sentences used in the training of the system (same text for training and testing). The third set of experiments, represented by Fig. 5, is speaker and text independent, and characterizes a scenario that best approximates a practical use of the LVDCSR system. A distribution of 75% and 25% of the speech database was used for training and testing in the experiments sets 1 and 2. The experiment set 3 used a distribution of 56.25% for training, 6.25% for testing and 37.5% of the database was not used.

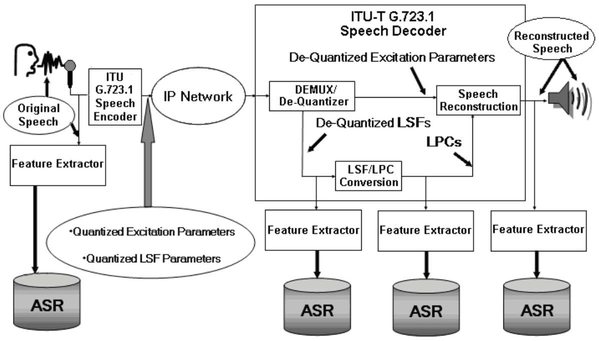


Fig. 2. Codec features extractors and ASR systems using the ITU-T G.723.1

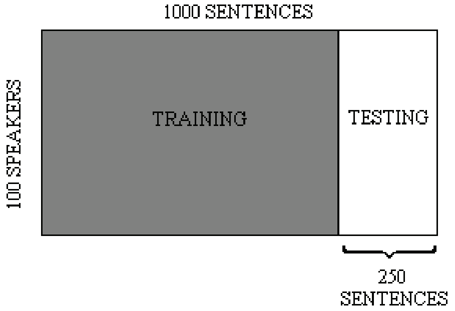


Fig. 3. Graphical representation of database used in Experiment set 1 (speaker dependent)

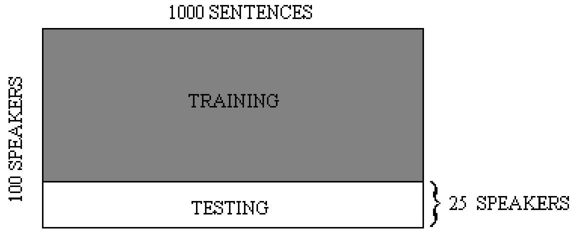


Fig. 4. Graphical representation of database used in Experiment set 2 (speaker independent with same text for training and testing)

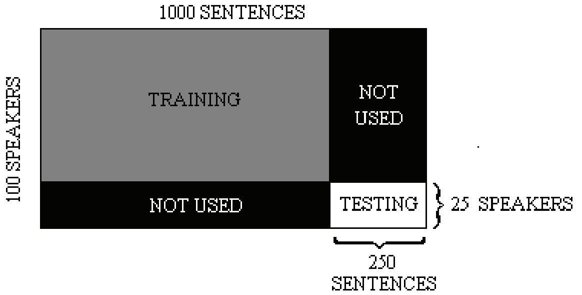


Fig. 5. Graphical representation of database used in Experiment set 3 (speaker and text independent)

To guarantee the statistical confidence of the results, cross-validation was employed in all experiments. Average Word Recognition Rates ($\overline{WRR}$), Standard Deviations (σ) and 95% Confidence Intervals are presented in the next section.

In all experiments in this work, the feature extractors generate one set of 10 parameters plus its first and second derivatives, representing a total of 30 recognition features. The Acoustic Model uses three states continuous observation HMMs (Hidden Markov Models) with a mixture of twenty Gaussians per state for phone modeling. Because silence is stationary, one state was used with the same number of Gaussians. They were implemented with the HTK (HMM Toolkit) software [8]. Inter- and intraword triphones are used as acoustic units. A Trigram language model was trained with the HTK (HMM Toolkit) software [8] with a lexicon of 60,080 words with perplexity of 307 obtained from 240,000 sentences extracted from a large text corpus of Ceten-Folha [15]. The Trigram language model was tested using the ATK (Application Toolkit for HTK) [16].

The system was simulated using a Sun V880 with 4 processors, 8Gb of RAM memory executing Solaris 10 operating system.

IV. Simulations Results

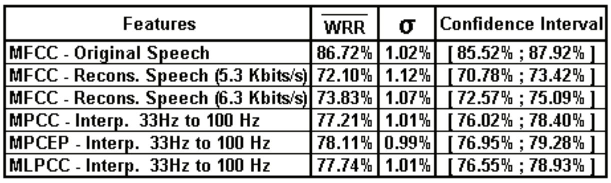
.

Performance results are given in three tables according to the experiments sets described in Section III. Table I shows the recognition results of experiment set 1 (the 100 speaker dependent scenario), Table II presents the recognition results of experiment set 2 (the speaker independent scenario with all the sentences used in the training and testing of the system) and Table III shows the recognition results of experiment set 3 (the speaker and sentences independent scenario). 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.

Comparing the results presented in Table I (speaker dependent scenario) with results presented in Table II (speaker independent scenario with all the sentences used in the training), we can see that the variability of the speaker yields a reduction of around 4% in the $\overline{WRR}$. It can also be seen that the performance drops near 6% from Table II (speaker independent scenario with all the sentences used in training) to Table III (speaker and sentences independent scenario). This shows that besides the 4% performance reduction due to speaker variability, an additional 6% occurs due to text variability. This is a result of different realization of the same triphone (different sentence contexts) during training and testing.

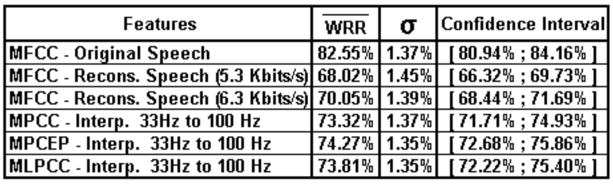
We have also obtained the recognition performance of the MFCC feature extracted from the Reconstructed Speech and Original Speech. Comparing the results of MFCC in theses two situations (Original vs Reconstructed), we can observe that the Reconstructed speech has high performance degradation (around 14%) as compared to the Original Speech, and is worse than the ones obtained with any of the recognition features MPCC, MPCEP and MLPCC in all experiments sets. The MFCC has also sensitivity of around of 2% by how the excitation is encoded and decoded (comparing row two – operation at 5.3 kbits/s – and row three – operation at 6.3 kbits/s – of each table). This shows that the MFCC is very sensitive to the encoding noise. The best results are obtained with the MPCEP recognition feature.

TABLE I
Recognition Accuracy in Experiment Set 1



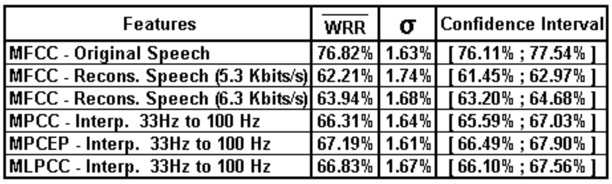
Features	$\overline{\text{WRR}}$	σ	Confidence Interval
MFCC - Original Speech	86.72%	1.02%	[85.52% ; 87.92%]
MFCC - Recons. Speech (5.3 Kbits/s)	72.10%	1.12%	[70.78% ; 73.42%]
MFCC - Recons. Speech (6.3 Kbits/s)	73.83%	1.07%	[72.57% ; 75.09%]
MPCC - Interp. 33Hz to 100 Hz	77.21%	1.01%	[76.02% ; 78.40%]
MPCEP - Interp. 33Hz to 100 Hz	78.11%	0.99%	[76.95% ; 79.28%]
MLPCC - Interp. 33Hz to 100 Hz	77.74%	1.01%	[76.55% ; 78.93%]

TABLE II.
Recognition Accuracy in Experiment Set 2



Features	$\overline{WRR}$	σ	Confidence Interval
MFCC - Original Speech	82.55%	1.37%	[80.94% ; 84.16%]
MFCC - Recons. Speech (5.3 Kbits/s)	68.02%	1.45%	[66.32% ; 69.73%]
MFCC - Recons. Speech (6.3 Kbits/s)	70.05%	1.39%	[68.44% ; 71.69%]
MPCC - Interp. 33Hz to 100 Hz	73.32%	1.37%	[71.71% ; 74.93%]
MPCEP - Interp. 33Hz to 100 Hz	74.27%	1.35%	[72.68% ; 75.86%]
MLPCC - Interp. 33Hz to 100 Hz	73.81%	1.35%	[72.22% ; 75.40%]

TABLE III
Recognition Accuracy in Experiment Set 3



Features	$\overline{WRR}$	σ	Confidence Interval
MFCC - Original Speech	76.82%	1.63%	[76.11% ; 77.54%]
MFCC - Recons. Speech (5.3 Kbits/s)	62.21%	1.74%	[61.45% ; 62.97%]
MFCC - Recons. Speech (6.3 Kbits/s)	63.94%	1.68%	[63.20% ; 64.68%]
MPCC - Interp. 33Hz to 100 Hz	66.31%	1.64%	[65.59% ; 67.03%]
MPCEP - Interp. 33Hz to 100 Hz	67.19%	1.61%	[66.49% ; 67.90%]
MLPCC - Interp. 33Hz to 100 Hz	66.83%	1.67%	[66.10% ; 67.56%]

V. Conclusion

In this paper, we have carried out several important experiments with Large Vocabulary Distributed Continuous Speech Recognition in Brazilian Portuguese. We have shown that only the independency of the speaker deteriorates in around 4% the recognition rate. An additional 6 % performance reduction is due to the use of different sentences during training and testing. We have also shown that the MFCC feature, which is obtained from the reconstructed speech, is highly sensitive to the encoding noise. The performance drops around 14%. We have observed that the MFCC of reconstructed speech has a sensitivity of around of 2% by how the excitation is encoded and decoded (comparing row two – operation at 5.3 kbits/s – and row three – operation at 6.3 kbits/s – of each table). The features obtained from the LSF or LPC parameters can provide much better recognition accuracy than the one obtained from the reconstructed speech (MFCC). The MPCEP feature is the best parameter to be used in an LVDCSR system employing the ITU-T G.723.1 codec. It yields the highest $\overline{WRR}$ with the lowest complexity.

Acknowledgment

The database developed to be used in this work was only possible due a big effort of Salt and Light (a non governmental organization in Brazil, that records books for blind people) and the support given by CNPQ (a governmental organization that supports research in Brazil).

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] V. F. S. Alencar and A. Alcaim, "Features Interpolation Domain for Distributed Speech Recognition and Performance for ITU-T G.723.1 CODEC", Proceedings of the ICSLP, Antwerp, BE, August 2007.

[4] ITU-T Recommendation G.723.1, "Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 Kbit/s," March 1996.

[5] Y. Ohshima, "Environmental Robustness in Speech Recognition using Physiologically-Motivated Signal Processing," PH. D. Thesis, Carnegie Mellon University, Pittsburgh, Pennsylvanya, December 1993.

[6] W. B. Kleijn and K. K. Paliwal, Speech Coding and Synthesis, Amsterdam, The Netherlands: Elsevier, 1995.

[7] 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.

[8] S. Young, et al., The HTK Book (for HTK Version 3.4), (http://htk.eng.cam.ac.uk/), December 2006.

[9] 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.

[10] S. K. Mitra, Digital Signal Processing: A Computer-Based Approach, McGraw-Hill International Editions, 1998.

[11] A. V. Oppenheim e D. H., Johnson, "Discrete Representation of Signals," Proc. IEEE, vol. 60, pp.681- 691, June 1972.

[12] M. Wölfel, J. McDonough, e A., Waibel, "Minimum Variance Distortionless Response on a Warped Frequency Scale," Eurospeech, Geneva, 2003.

[13] F. S. Gurgen, S. Sagayama, e S. Furui, "Line Spectrum Frequency-Based Distance Measures for Speech Recognition," pp.521-524, Proc. ICSLP, Kobe, Japan, November 1990.

[14] R. J. R. Cirigliano, et al., "Um Conjunto de 1000 Frases Foneticamente Balanceadas para o Português Brasileiro obtido utilizando a abordagem de Algoritmos Genéticos", XXII Simpósio Brasileiro de Telecomunicações – SBrT, 2005.

[15] "Corpus de Extractus de Textos Eletrônicos Nilc/ Folha de São Paulo (Ceten-Folha)", (http://acdc.linguateca.pt/cetenfolha/), 14 November 2005.

[16] S. Young, An Application Toolkit for HTK (ATK 1.6), (http://htk.eng.cam.ac.uk/develop/atk.shtml), June 2007.

Digital filter interpolation of decoded LSFs for distributed continuous speech recognition

V.F.S. de Alencar and A. Alcaim

A digital filter interpolation of decoded line spectral frequencies (LSFs) that significantly outperforms linear interpolation for large vocabulary distributed continuous speech recognition systems is presented. Experiments were conducted using linear predictive coding (LPC) and LSF-derived speech recognition features, CDHMM acoustic models, triphone units and trigram language models for Brazilian Portuguese.

Introduction: The last few years have witnessed a considerable change in the way speech is carried over digital communication networks. On the other hand, the rapid growth of both mobile and IP networks motivated the integration of automatic speech recognition (ASR) technologies into these networks. Owing to the high complexity of ASR, distributed recognition systems, where acoustic parameters are extracted at the user terminal and recognition is performed at the remote server, are used. In this Letter, we focus on a user terminal where speech is encoded by the ITU-T G.723.1 codec [1]. It is one of the most commonly used codecs for VoIP transmission, owing to its high compression rates (5.3 or 6.3 kbit/s) and the quality of the decoded speech.

The ITU-T G.723.1 utilises linear predictive coding (LPC) algorithms based on a speech production model. In this model, an excitation signal is applied to an all-pole filter (characterised by the LPC parameters), that represents the spectral envelope information of the speech signals. Usually, the LPC parameters are transformed to line spectral frequencies (LSF), owing to the attractive properties of the latter to the quantisation interpolation procedures. It is also known that extracting speech recognition features from the parameters of a speech coder provides better gnition performance than obtaining the features from the decoded/ nstructed signal [2]. However, the parameters of a speech coder are the most adequate ones for the remote recognition system. For this on, different codec parameter transformations have to be considered s to improve recognition accuracy.

nother important remark is that, for satisfactory operation of the R system, the recognition features have to be obtained at a high (typically 100 Hz). However, speech coders for mobile telephony IP networks generate their parameters at lower rates (e.g. 33 Hz). recent study on the efficiency of recognition features for distributed ech recognition [3], it was shown that low rates significantly degrade performance of the recogniser. Usually the linear interpolation in the domain [4] is used to solve this problem. In this Letter we propose a tal filter interpolation technique that results in better speech recogon performance.

xperimental procedures: The new database used in this Letter was designed on the basis of a phonetically balanced set of 1000 sentences for Brazilian Portuguese [5]. The speech database comprises 50 male speakers and 50 female speakers, each one repeating once all the 1000 sentences (3528 words). The database was recorded in a studio with 16 kHz sample rate and 16 bits per sample with a bandwith 50–7000 Hz. This database was filtered and down-sampled to match the ITU-T G723.1 [1] requirements.

The experiments were carried out in a speaker and text independent mode, and characterise a scenario that best approximates a practical use of distributed speech recognition system. We used a distribution of 56.25% of the database for training (75 speakers, each uttering 750 sentences), 6.25% for testing (25 different speakers, each uttering 250 different sentences) and 37.5% of the database was not used. To guarantee the statistical confidence of the results, cross-validation was employed in all experiments. Performance was measured in terms of average word recognition rates ($\overline{WRR}$), standard deviations (σ) and 95% confidence intervals.

The feature extractors generate one set of 10 parameters plus its first and second derivatives, representing a total of 30 recognition features. The acoustic model uses three states continuous observation hidden Markov models (HMMs) with a mixture of 20 Gaussians per state for phone modelling. Because silence is stationary, one state was used with the same number of Gaussians. Inter- and intraword triphones are used as acoustic units. A trigram language model for Brazilian Portuguese was trained with a lexicon of 60 080 words with perplexity of 307 obtained from 240 000 sentences extracted from a large text corpus of Ceten-Folha [6].

In this Letter, we consider the MEL scale features obtained from LPC (the MLPCC [7]) and from LSF (the MPCC [8] and MPCEP [2]), since they provide much better performance than those achieved with the linear scale features (LPCC, PCC and PCEP) [3]. For comparative purposes we have also obtained the mel-frequency cepstral coefficients (MFCC) features from the original speech and from voice reconstructed with the ITU-T G.723.1 codec at the two different operation rates (6.3 or 5.3 kbit/s) [1].

LSFs interpolation: The linear interpolation is a technique usually employed in a variety of applications to obtain a signal at a higher rate. In particular, it has been used in distributed speech recognition systems to interpolate the decoded LSFs from the bitstream of the IS-641 speech coder in [9] and from the bitstream of the ITU-T G.723.1 codec in [4]. In the latter case the LSFs are obtained at the higher rate of 100 Hz from the ITU-T G.723.1 codec which operates at 33 Hz (one set of parameters every 30 ms) [4, 9].

In this Letter, we propose an interpolation technique that is designed using an up-sampler followed by a lowpass digital filter $H(z)$. The up-sampler with factor $r > 1$ (where r is the interpolation factor that in this case is 3), inserts $r - 1$ equidistant zero-valued samples between two consecutive samples. The lowpass filter $H(z)$ eliminates the insertion of images (in this case two images) of the original spectrum compressed by a factor r. It should be noted that in this interpolation application, one important requirement is to ensure that the sequence of input samples (in our case, the decoded LSFs) are not changed at the output. The input signal is assumed to be finite energy and band limited to the frequency range $0 \leq \omega \leq \alpha$, where α must be equal or smaller than 0.5 (in our case, α is assumed to be equal to 0.5). The filter $H(z)$ is a symmetric FIR filter that determines the missing samples by minimising the mean square errors using the orthogonality principle [10]. The length of the FIR filter used in this Letter is $2rL + 1$, where L is an integer that determines the length of the filter (in this Letter L was made equal 4). To reduce the complexity of directly computing the coefficients of this filter of length $2rL + 1 = 25$, we used the techniques proposed in [10] to design optimum interpolators with lower complexity and consuming less hardware. The procedure corresponds to replacing the design of a filter of length 25 by $r = 3$ subfilters of length $2L + 1 = 9$.

Simulation results: Table 1 (rows 4 to 6) shows the recognition results for the interpolation of the recognition features using the ITU-T G723.1 speech codec. It should be noted that in each test case the model parameters are trained with the same type of features and interpolation, i.e. training and testing are matched. The interpolation is carried out in the LSF domain, the MPCEP and MPCC features are directly obtained from the interpolated LSFs. On the other hand, the MLPCC features are obtained from the LPC parameters which, in turn, are generated from the interpolated LSFs.

Table 1: Recognition accuracy for linear and digital filter interpolation techniques

Features	Linear interpolation			Digital filter interpolation		
	$\overline{WRR}$ (%)	σ(%)	Confidence interval (%)	$\overline{WRR}$ (%)	σ(%)	Confidence interval (%)
MFCC–original speech	76.82	1.63	[76.11; 77.54]	76.82	1.63	[76.11; 77.54]
MFCC–reconst. (5.3 kbits/s)	62.21	1.74	[61.45; 62.97]	62.21	1.74	[61.45; 62.97]
MFCC–reconst. (6.3 kbits/s)	63.94	1.68	[63.20; 64.68]	63.94	1.68	[63.20; 64.68]
MPCC–interp. 33–100 Hz	66.31	1.64	[65.59; 67.03]	70.21	1.59	[69.51; 70.91]
MPCEP–interp. 33–100 Hz	67.19	1.61	[66.49; 67.90]	71.32	1.57	[70.63; 72.01]
MLPCC–interp. 33–100 Hz	66.83	1.67	[66.10; 67.56]	70.85	1.60	[70.15; 71.55]

Comparing the results presented in Table 1 we can see that the digital filter interpolation of decoded LSFs can provide remarkable recognition improvement in the distributed systems. It yields an average performance gain of around 4% when compared to the usual linear interpolation procedure. The best word recognition rate was achieved by the MPCEP feature (71.32% recognition rate), using the digital filter interpolation of the decoded LSFs. From the Table it can also be observed that when the MFCC feature is used, the reconstructed speech presents a high performance degradation compared to the original speech. Moreover, the MFCC of reconstructed speech is worse than that obtained with any of the LSF- and LPC-derived recognition features (MPCC, MPCEP, MLPCC) in all experiments.

Conclusion: Speech coders for mobile telephony and IP networks (e.g. the ITU-T G.723.1) generate their parameters at a rate usually lower than those required for speech recognition. Therefore, in distributed speech recognition it is paramount to interpolate them. In this Letter, we have investigated the use of a digital filter interpolation in a scenario of large vocabulary distributed continuous speech recognition in Brazilian Portuguese. We have shown that, when compared to linear interpolation, the digital filter interpolation of the decoded LSFs remarkably improved the performance of all the recognition features obtained from the bitstream of the ITU-T G.723.1 codec. The average word recognition rate gain was approximately 4% in all situations where recognition features were derived from the decoded LSFs or the LPC parameters.

Acknowledgments: The database developed for use in this work was only possible because of much effort by Salt and Light (a non-governmental organisation in Brazil, that records books for blind people) and the support given by CNPQ (a governmental organisation that supports research in Brazil).

10 June 2008
Electronics Letters online no: 20089650
doi: 10.1049/el:20089650

V.F.S. de Alencar and A. Alcaim (*Centre for Telecommunications Studies of the Catholic University (CETUC) PUC-RIO, Rio de Janeiro 22453-900, Brazil*)

E-mail: vladimir@cetuc.puc-rio.br

References

1 ITU-T Recommendation G.723.1, 'Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 Kbit/s,' March 1996

2 Choi, H.S., Kim, H.K., and Lee, H.S.: 'Speech recognition using quantized LSP parameters and their transformations in digital communication', *Speech Commun.*, 2000, **30**, pp. 223–233

3 Alencar, V.F.S., and Alcaim, A.: 'Transformations of LPC and LSF parameters to speech recognition features'. Proceedings of ICAPR, Bath, UK, August 2005

4 Alencar, V.F.S., and Alcaim, A.: 'Features interpolation domain for distributed speech recognition and performance for ITU-T G.723.1 CODEC'. Proc. of ICSLP, Antwerp, Belgium, August 2007

5 Cirigliano, R.J.R., *et al.*: 'Um Conjunto de 1000 Frases Foneticamente Balanceadas para o Português Brasileiro Obtido Utilizando a Abordagem de Algoritmos Genéticos'. XXII Simpósio Brasileiro de Telecomunicações – SBrT, 2005

6 'Corpus de Extractus de Textos Eletrônicos Nilc/Folha de São Paulo (Ceten-Folha)', (http://acdc.linguateca.pt/cetenfolha/), 14 November 2005

7 Ohshima, Y.: 'Environmental robustness in speech recognition using physiologically-motivated signal processing', Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, December 1993

8 Kleijn, W.B., and Paliwal, K.K.: 'Speech coding and synthesis' (Elsevier, Amsterdam, The Netherlands, 1995)

9 Kim, H.K., and Cox, R.V.: 'A bitstream-based front-end for wireless speech recognition on IS-136 communications system', *IEEE Trans. Speech Audio Process.*, 2001, **9**, pp. 558–568

10 Oetken, G., Parks, T.W., and Schüssler, H.W.: 'New results in design of digital interpolators', *IEEE Trans. Accoust. Speech Signal Process.*, 1975, **ASSP-23**, pp. 301–309

On the Performance of ITU-T G.723.1 and AMR-NB Codecs for Large Vocabulary Distributed Speech Recognition in Brazilian Portuguese

Vladimir Fabregas Surigué de Alencar and Abraham Alcaim
CETUC – PUC/RIO – 22453-900, Rio de Janeiro – Brazil
vladimir@cetuc.puc-rio.br, alcaim@cetuc.puc-rio.br

Abstract

In this paper, we present the accuracy for large vocabulary distributed continuous speech recognition systems over ITU-T G.723.1 and AMR-NB speech codecs. Experiments were conducted using LPC and LSF-derived speech recognition features, CDHMM acoustic models, triphone units and trigram language models for the Brazilian Portuguese.

1. Introduction

The last few years have witnessed a considerable change in the way speech is carried over digital communications networks. On the other hand, the rapid growth of both mobile and IP networks motivated the integration of Automatic Speech Recognition (ASR) technologies into these networks. Due to the high complexity of ASR, distributed recognition systems, where acoustic parameters are extracted at the user terminal and recognition is performed at the remote server, are used. In this paper, we have focused on a user terminal where speech is encoded by the ITU-T G.723.1 [1] or AMR-NB [2] codec. The ITU-T G723.1 is one of the most commonly used codecs for VoIP transmission, due to its high compression rates (5.3 or 6.3 kbit/sec) and to the quality of the decoded speech. The AMR-NB is the standard codec for the Global System for Mobile Communications (GSM).

Both ITU-T G.723.1 and AMR-NB utilize 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 signals. 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 extracting speech recognition features from the parameters of a speech coder provides better recognition performance than obtaining the features from the decoded/reconstructed signal [3]. 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., 33 or 50 Hz). In a recent study on the efficiency of recognition features for distributed speech recognition [4], it was shown that reduction of rate significantly degrades the performance of the recognizer. Usually the linear interpolation in the LSF domain [5] is used to solve this problem. However, in this paper we will use a digital filter interpolation technique that results in a better speech recognition performance [6] when compared with linear interpolation.

This paper is organized as follows. Section 2 and 3 give a brief overview of the recognition features and codecs used in this paper. Section 4 describes the LSFs interpolation procedure, Section 5 the database and experiments and Section 6 the simulation results. Section 7 concludes the paper.

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) [7].

The Line Spectral Frequencies (LSFs) are often used in speech coders due to their high coding efficiency and attractive interpolation properties [8]. 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) [9], MPCC (Mel-Frequency PCC) [9], PCEP (Pseudo-Cepstrum) [3] and MPCEP (Mel-Frequency PCEP) [3]. 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) [4]. The MFCC (Mel-Frequency Cepstral Coefficients) features [10]-[11] will be also obtained from voice reconstructed with the ITU-T G.723.1 codec and AMR-NB codecs. The difference between the rates of ITU-T G723.1 and AMR-NB affects the performance obtained with the recognition parameters and will be observed from the simulation results.

The detailed equations describing the MLPCC, MPCC and MPCEP features are given in [4].

3. The ITU-T G.723.1 and AMR-NB Codecs

The ITU-T G.723.1 codec operates at the following bit-rates: 5.3 and 6.3 kbits/s. At the 6.3 kbits/s the ITU-T G.723.1 employs a Multipulse Maximum Likelihood Quantization (MP-MLQ) [12]. At the 5.3 kbits/s the ITU-T G.723.1 uses an Algebraic-Code-Excited Linear-Prediction (ACELP) [13]. It transmits two types of information, to be used at the receiver for synthesizing the speech signal: LSF (Line Spectral Frequencies), that represents the frequency response of the synthesis filter, and the excitation signal to the synthesis filter.

The coder is based on the principles of linear prediction analysis-by-synthesis coding and attempts to minimize a perceptually weighted error signal. The encoder operates on blocks (frames) of 240 samples each. That is equal to 30 msec at an 8 kHz sampling rate. Each block is first high pass filtered to remove the DC component and then divided into four subframes of 60 samples each. For every subframe, a 10th order Linear Prediction Coder (LPC) filter is computed using the unprocessed input signal. The LPC filter for the last subframe is converted to LSF and quantized using a Predictive Split Vector Quantizer (PSVQ). The LSF parameters are encoded with 24 bits/frame for both coded rates.

The ITU-T G.723.1 encoder is dedicated to compress the voice signals with bandwidth up to 4 kHz efficiently and to deliver an encoded data stream with a very low binary rate and a good quality of transmitted speech – typical applications being encoding of the vocal signal for video conferences via GSTN (General Switch Telecommunication Network) and Voice over IP.

The AMR-NB codec operates at the following bit-rates: 4.75, 5.15, 5.9, 6.7, 7.4, 7.95, 10.2 and 12.2 kbits/s. The AMR-NB is a ACELP type codec.

The coder operates on speech frames of 20 ms corresponding to 160 samples at the sampling frequency of 8000 sample/s. LP analysis is performed twice per frame for the 12.2 kbits/s mode and once for the other modes. For the 12.2 kbits/s mode, the two sets of LP parameters are converted to LSF and jointly quantized using Split Matrix Quantization (SMQ) with 38 bits/frame. For the other modes, the single set of LP parameters is converted to Linear Spectral Frequencies (LSF) and vector quantized using Split Vector Quantization (SVQ). At the 10.2, 7.4, 6.7 and 5.9 kbits/s the LSFs are quantized with 26 bits/frame. At the operating rate 7.95 the LSF are encoded with 27 bits/frame. At the 5.15 and 4.75 kbits/s the LSFs are quantized with 23 bits/frame.

The different bit-rates of this codec are commonly referred to as modes. The original idea has been that the modes could automatically be altered to ensure the best possible bandwidth sharing between the speech coder and the channel coder. In case of poor channel conditions, to maximize the error protection, the low bit-rate modes are selected, and in the presence of a good channel, the higher bit-rates are used. In this way, the perceptual speech quality can be kept at the highest possible level.

The standardisation in 1999 [14] of AMR-NB as the speech codec of GSM represented a big improvement of voice quality for this mobile network. The AMR-NB codec was also adopted in 1999 by 3GPP as the default speech codec for the WCDMA 3G system. The AMR codec was jointly developed by Ericsson, Nokia and Siemens.

4. LSFs Interpolation

The Linear Interpolation is a technique usually employed in a variety of applications to obtain a signal at a higher rate. In particular, it has been used in distributed speech recognition systems to interpolate the decoded LSFs from the bitstream of the IS-641 speech coder in [15] and from the bitstream of the ITU-T G.723.1 codec in [5]. In the later case the LSFs are obtained at the higher rate of 100Hz from the ITU-T G.723.1 codec which operates at 33 Hz (one set of parameters every 30 ms) [5] [15].

In [6] a new interpolation technique was proposed that outperforms the Linear Interpolation and for this reason will be the one to be used in this paper. This new interpolation technique is designed using an up-sampler followed by a lowpass digital filter $H(z)$. The up-sampler with factor $r > 1$ (where r is the interpolation factor that in the case of ITU-T G.723.1 is 3 and in the case of AMR-NB is 2), inserts $r-1$ equidistant zero-valued samples between two consecutive samples. The lowpass filter $H(z)$ eliminates the insertion of images (in this case two images) of the original spectrum compressed by a factor r. It should be noted that in this interpolation application, one important requirement is to ensure that the sequence of input samples (in our case, the decoded LSFs) are not changed at the output. The input signal is assumed to be finite energy and band limited to the frequency range $0 \leq \omega \leq \alpha$, where α must be equal or smaller than 0.5 (in our case, α is assumed to be equal to 0.5). The filter $H(z)$ is a symmetric FIR filter that determines the missing samples by minimizing the mean square errors using the orthogonality principle [16]. The length of the FIR filter used in this letter is $2rL+1$, where L is an integer that determines the length of the filter (in this paper L was made equal 4). In order to reduce the complexity of directly computing the coefficients of this filter of length $2rL+1$ (equal 25 for ITU-T G.723.1 and 17 for AMR-NB), we have used the techniques proposed in [16] to design optimum interpolators with lower complexity and less hardware consuming. The procedure corresponds to replacing the design of a filter of length $2rL+1$(equal 25 for ITU-T G.723.1 and 17 for AMR-NB) by r (equal 3 for ITU-T G.723.1 and 2 for AMR-NB) subfilters of length $2L+1$ (equal 9 for ITU-T G.723.1 and 9 for AMR-NB).

5. Description of the Database and Experiments

The database used in this paper was designed on the basis of a phonetically balanced set of 1000 sentences for the Brazilian Portuguese [17]. The speech database is composed of 50 male speakers and 50 female speakers, each one repeating once all the 1000 sentences (3528 words). The database was recorded in a studio at a 16 kHz sample rate and 16 bits per sample with a bandwith from 50 – 7000 Hz. This database was filtered and down sampled to match the ITU-T G723.1 [1] and AMR-NB [2] requirements.

The experiments, were carried out in a speaker and text independent mode, and characterize a scenario that best approximates a practical use of distributed speech recognition system. We have used a distribution of 56.25% of the database for training (75 speakers, each uttering 750 sentences), 6.25% for testing (25 different speakers, each uttering 250 different sentences) and 37.5% of the database was not used. To guarantee the statistical confidence of the results, cross-validation was employed in all experiments. Performance was measured in terms of Average Word Recognition Rates ($\overline{WRR}$), Standard Deviations (σ) and 95% Confidence Intervals.

The feature extractors generate one set of 10 parameters plus its first and second derivatives, representing a total of 30 recognition features. The Acoustic Model uses three states continuous observation HMMs (Hidden Markov Models) with a mixture of twenty Gaussians per state for phone modeling. Because silence is stationary, one state was used with the same number of Gaussians. Inter- and intraword triphones are used as acoustic units. A Trigram language model for the Brazilian Portuguese was trained with a lexicon of 60,080 words with perplexity of 307 obtained from 240,000 sentences extracted from a large text corpus of Ceten-Folha [18].

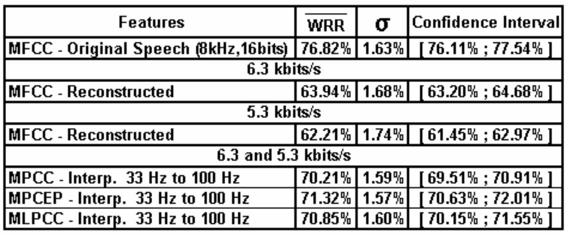
6. Simulation Results

Performance results are given in two tables, Table.1 shows the recognition results for the ITU-T G723.1 and Table.2 shows the recognition results for the AMR-NB. 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.

It is also important to remember that the AMR-NB operating at 12.2 kbits/s generate LSFs at 100 Hz avoiding the need to interpolate de LSF for this mode of the codec (for the other speech codec rates LSF are generate at 50 Hz and need to be interpolated to achieve the 100 Hz).

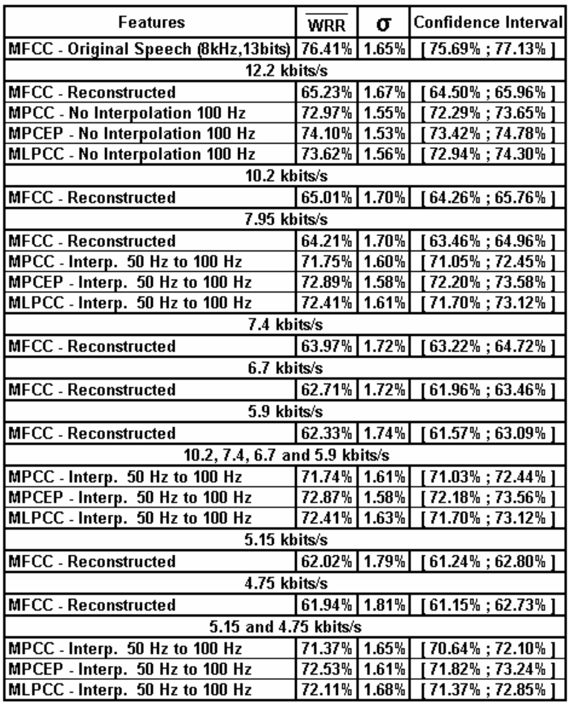
We have also obtained the recognition performance of the MFCC feature extracted from the Reconstructed Speech and Original Speech. Comparing the results of MFCC in theses two situations (Original vs Reconstructed), we can observe that the Reconstructed speech has high performance degradation (between 11% and 15%) as compared to the Original Speech, and is worse than the ones obtained with any of the recognition features MPCC, MPCEP and MLPCC in all experiments sets. This shows that the MFCC is very sensitive to the encoding noise. The best results are obtained with the MPCEP recognition feature.

Table 1. Recognition accuracy in ITU-T G.723.1



Features	$\overline{WRR}$	σ	Confidence Interval
MFCC - Original Speech (8kHz,16bits)	76.82%	1.63%	[76.11% ; 77.54%]
6.3 kbits/s			
MFCC - Reconstructed	63.94%	1.68%	[63.20% ; 64.68%]
5.3 kbits/s			
MFCC - Reconstructed	62.21%	1.74%	[61.45% ; 62.97%]
6.3 and 5.3 kbits/s			
MPCC - Interp. 33 Hz to 100 Hz	70.21%	1.59%	[69.51% ; 70.91%]
MPCEP - Interp. 33 Hz to 100 Hz	71.32%	1.57%	[70.63% ; 72.01%]
MLPCC - Interp. 33 Hz to 100 Hz	70.85%	1.60%	[70.15% ; 71.55%]

Table 2. Recognition accuracy in AMR-NB



Features	$\overline{WRR}$	σ	Confidence Interval
MFCC - Original Speech (8kHz,13bits)	76.41%	1.65%	[75.69% ; 77.13%]
12.2 kbits/s			
MFCC - Reconstructed	65.23%	1.67%	[64.50% ; 65.96%]
MPCC - No Interpolation 100 Hz	72.97%	1.55%	[72.29% ; 73.65%]
MPCEP - No Interpolation 100 Hz	74.10%	1.53%	[73.42% ; 74.78%]
MLPCC - No Interpolation 100 Hz	73.62%	1.56%	[72.94% ; 74.30%]
10.2 kbits/s			
MFCC - Reconstructed	65.01%	1.70%	[64.26% ; 65.76%]
7.95 kbits/s			
MFCC - Reconstructed	64.21%	1.70%	[63.46% ; 64.96%]
MPCC - Interp. 50 Hz to 100 Hz	71.75%	1.60%	[71.05% ; 72.45%]
MPCEP - Interp. 50 Hz to 100 Hz	72.89%	1.58%	[72.20% ; 73.58%]
MLPCC - Interp. 50 Hz to 100 Hz	72.41%	1.61%	[71.70% ; 73.12%]
7.4 kbits/s			
MFCC - Reconstructed	63.97%	1.72%	[63.22% ; 64.72%]
6.7 kbits/s			
MFCC - Reconstructed	62.71%	1.72%	[61.96% ; 63.46%]
5.9 kbits/s			
MFCC - Reconstructed	62.33%	1.74%	[61.57% ; 63.09%]
10.2, 7.4, 6.7 and 5.9 kbits/s			
MPCC - Interp. 50 Hz to 100 Hz	71.74%	1.61%	[71.03% ; 72.44%]
MPCEP - Interp. 50 Hz to 100 Hz	72.87%	1.58%	[72.18% ; 73.56%]
MLPCC - Interp. 50 Hz to 100 Hz	72.41%	1.63%	[71.70% ; 73.12%]
5.15 kbits/s			
MFCC - Reconstructed	62.02%	1.79%	[61.24% ; 62.80%]
4.75 kbits/s			
MFCC - Reconstructed	61.94%	1.81%	[61.15% ; 62.73%]
5.15 and 4.75 kbits/s			
MPCC - Interp. 50 Hz to 100 Hz	71.37%	1.65%	[70.64% ; 72.10%]
MPCEP - Interp. 50 Hz to 100 Hz	72.53%	1.61%	[71.82% ; 73.24%]
MLPCC - Interp. 50 Hz to 100 Hz	72.11%	1.68%	[71.37% ; 72.85%]

Now comparing the ITU-T G723.1 at the rates 6.3 and 5.3 kbits/s with the AMR-NB codec at similar rates (6.7 and 5.9 kbits/s) the latter can provide a 1.55% of $\overline{WRR}$ improvement. It is also very significant to note that the 95% confidence interval in these two cases do not overlap.It is also important to remark that the AMR-NB at lower rates (5.15 and 4.75 kbits/s) outperform the ITU-T G.723.1 codec operating at 6.3 and 5.3 kbits/s. Moreover, their 95% confidence interval again do not overlap.

Finally, should be noted that the LSFs in the AMR-NB are encoded at a higher bit rate than the one used by the ITU-T G.723.1.

7. Concluding Remarks

In this paper, we have carried out several important experiments with Large Vocabulary Distributed Continuous Speech Recognition in Brazilian Portuguese. We have also shown that the MFCC feature, which is obtained from the reconstructed speech, is highly sensitive to the encoding noise. The performance drops between 11% and 15%. The features obtained from the LSF or LPC parameters can provide much better recognition accuracy than the one obtained from the reconstructed speech (MFCC). The MPCEP feature is the best parameter to be used in an LVDCSR system employing the ITU-T G.723.1 or AMR-NB codec. It yields the highest $\overline{WRR}$ with the lowest complexity. In addition, the AMR-NB operating at a lower bit rate overperforms the ITU-T G.723.1 codec without overlapping their confidence interval.

8. References

[1] ITU-T Recommendation G.723.1, "Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 Kbit/s," March 1996

[2] 3GPP TS 26.071 V6.0.0, "Mandatory speech CODEC speech processing functions, AMR speech CODEC - General description," December, 2004.

[3] 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

[4] 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

[5] V. F. S. Alencar and A. Alcaim, "Features Interpolation Domain for Distributed Speech Recognition and Performance for ITU-T G.723.1 CODEC", Proceedings of the ICSLP, Antwerp, BE, August 2007

[6] V. F. S. Alencar and A. Alcaim, "Digital Filter Interpolation of Decoded LSFs for Distributed Continuous Speech Recognition", Electronics Letters, vol.44, issue:17, pp.1039-1040, August 2008.

[7] Y. Ohshima, "Environmental Robustness in Speech Recognition using Physiologically-Motivated Signal Processing," PH. D. Thesis, Carnegie Mellon University, Pittsburgh, Pennsylvanya, December 1993

[8] W. B. Kleijn and K. K. Paliwal, Speech Coding and Synthesis, Amsterdam, The Netherlands: Elsevier, 1995

[9] 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.

[10] S. Young, et al., The HTK Book (for HTK Version 3.4), (http://htk.eng.cam.ac.uk/), December 2006.

[11] 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.

[12] B.S. Atal, and J.R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," Proceedings of ICASSP, pp. 614–617, 1982.

[13] R. Salami, C. Laflamme, J. P. Adoul, and D. Massaloux, "A toll quality 8 kb/s speech codec for the personal communications system (PCS)," IEEE Trans. Vehicular Technol., vol. 43, pp. 808–816, Aug. 1994.

[14] K. Järvinen, "Standardisation of the Adaptive Multi-rate Codec," European Signal Processing Conference (EUSIPCO), Tampere, Finland, 4–8 Sept. 2000.

[15] H. K. Kim and R. V. Cox, "A Bitstream-Based Front-End for Wireless Speech Recognition on IS-136 Communications System," IEEE Trans. On Speech and Audio Processing, vol. 9, pp. 558-568, July 2001.

[16] G. Oetken, T. W. Parks and H. W. Schüssler, "New results in design of digital interpolators," IEEE Trans. On Accoustics, Speech & Signal Processing, ASSP-23:301-309, June 1975

[17] R. J. R. Cirigliano, et al., "Um Conjunto de 1000 Frases Foneticamente Balanceadas para o Português Brasileiro obtido utilizando a abordagem de Algoritmos Genéticos", XXII Simpósio Brasileiro de Telecomunicações – SBrT, 2005

[18] "Corpus de Extractus de Textos Eletrônicos Nilc/ Folha de São Paulo (Ceten-Folha)", (http://acdc.linguateca.pt/cetenfolha/), 14 November 2005

Acknowledgment

The database developed to be used in this work was only possible due a big effort of Salt and Light (a non governmental organization in Brazil, that records books for blind people) and the support given by CNPQ (a governmental organization that supports research in Brazil).

Recuperação de Pacotes Perdidos em Sistemas de Reconhecimento de Voz Distribuído usando Redes Neurais

Vladimir F. S. de Alencar e Abraham Alcaim

***Resumo*—Este artigo propõe uma nova técnica de reconstrução de pacotes perdidos em rajadas em sistemas de reconhecimento de voz distribuído com amplo vocabulário utilizando os codificadores de voz ITU-T G.723.1 e AMR-NB. A nova técnica, que é baseada em Redes Neurais, explora o conhecimento do sinal sem inserir um atraso significante. Experimentos foram conduzidos utilizando o atributo de reconhecimento de voz derivado de LSF (MPCEP), modelos acústicos CDHMM (Continuous Density HMM), unidades trifone e modelos de linguagem trigrama para o Português Brasileiro. Resultados de simulação mostraram que a técnica proposta supera o desempenho de reconhecimento quando comparada com as técnicas de Inserção de Zeros e a Interpolação Linear.**

***Palavras-Chave*— Redes Neurais, Reconhecimento de Voz Distribuído, ITU-T G.723.1, AMR-NB, LSF, LPC, HMM.**

***Abstract*—In this Paper, we propose a novel technique to reconstruct burst-like lost packets in large vocabulary distributed continuous speech recognition systems operating with ITU-T G.723.1 and AMR-NB speech codecs. The new technique, which is based on Neural Networks, takes advantage of the knowledge of the signal without inserting any significant delay. Experiments were conducted using an LSF-derived speech recognition feature (MPCEP), CDHMM (Continuous Density HMM) acoustic models, triphone units and trigram language models for the Brazilian Portuguese. Simulation results show that the proposed technique improves recognition performance as compared to Zero Insertion and Linear Interpolation schemes.**

***Index Terms*— Neural Networks, Distributed Speech Recognition, ITU-T G.723.1, AMR-NB, LSF, LPC, HMM.**

I. INTRODUÇÃO

O desenvolvimento tecnológico do mundo atual tem estimulado a demanda cada vez maior por máquinas inteligentes. Dentro desse panorama, a área de reconhecimento automático de voz (RAV) é uma das que tem despertado maior interesse, apesar da grande complexidade envolvida em termos de projeto e de operação. Esse interesse crescente tem sido evidente tanto no âmbito das indústrias como dos centros de pesquisa no mundo inteiro. Tendo em vista o crescimento gigantesco da Internet e dos sistemas de comunicações móveis celulares, as aplicações de processamento de voz nesses meios têm despertado interesses cada vez maiores. Em particular, um problema importante nessa área diz respeito ao reconhecimento de voz em um sistema servidor, a partir de parâmetros acústicos calculados e quantizados no terminal do usuário. O servidor reconhece a voz de acordo com uma aplicação específica e envia de volta, ao usuário, informações relativas à ação tomada a partir do reconhecimento de voz.

Devido à alta complexidade computacional e à grande quantidade de memória requerida em sistemas de RAV, se torna muito atraente a opção por sistemas de reconhecimento de voz distribuídos. Em sistemas desse tipo, o processamento é distribuído entre o terminal do usuário (telefone celular, computador pessoal) e o terminal de recepção em uma rede de comunicações (estação base em redes de telefonia móvel, servidor central em redes IP). Por esse motivo, para o desenvolvimento de sistemas voltados a estas redes é necessário conhecer os codificadores de voz utilizados nas mesmas.

Neste artigo, nos baseamos em um terminal usuário onde a voz fosse codificada pelo codec ITU-T G.723.1 [1] ou AMR-NB [2]. O ITU-T G.723.1 é um dos codecs mais amplamente utilizados para a transmissão de voz sobre IP (VoIP), devido a suas taxas elevadas da compressão (5,3 ou 6,3 kbit/s) e à qualidade da voz decodificada. O AMR-NB é o codec padrão para o Sistema Global para as Comunicações Móveis (GSM).

Os Codificadores ITU-T G.723.1 e AMR-NB utilizam os algoritmos LPC (Linear Predictive Coding) baseados em um modelo da produção da voz. Neste modelo, um sinal da excitação é aplicado a um filtro só de pólos (caracterizado pelos parâmetros LPC), o qual representa a informação espectral do envelope do sinal de voz. Geralmente, os parâmetros do LPC são transformados em LSF (Linear Spectral Frequencies), devido às propriedades atrativas do último para os procedimentos de quantização e de interpolação. Sabe-se também que extrair os atributos de reconhecimento dos parâmetros de um codificador de voz fornece um desempenho melhor de reconhecimento do que se obtendo os atributos do sinal decodificado/reconstruído [3]. Entretanto, os parâmetros dos codificadores de voz não são os mais adequados para o sistema de reconhecimento remoto. Por esta razão, diferentes transformações dos parâmetros dos codecs foram consideradas a fim melhorar o desempenho de

Vladimir F. S. de Alencar e Abraham Alcaim, CETUC, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brasil, E-mails: vladimir@cetuc.puc-rio.br, alcaim@cetuc.puc-rio.br.

reconhecimento. Neste artigo, nós consideraremos somente o atributo MPCEP (Mel Frequency Pseudo Cepstrum), pois o mesmo demonstrou em [4] que fornece um desempenho melhor, com uma complexidade menor, do que outros atributos de reconhecimento obtidos dos parâmetros do codec.

Uma outra observação importante é que para o funcionamento satisfatório dos sistemas RAV, os atributos de reconhecimento têm que ser obtidos em uma taxa elevada (tipicamente 100 Hz). Entretanto, os codificadores de voz para a telefonia móvel e redes IP geram seus parâmetros em taxas mais baixas (por exemplo, 33 ou 50 Hz). A Interpolação Linear no domínio das LSF [5] é usada geralmente para resolver este problema. Entretanto, neste artigo, nós usaremos uma técnica de interpolação com filtro digital (também no domínio das LSFs) que apresenta um desempenho melhor no reconhecimento da voz [6] quando comparada com a interpolação linear.

O problema de perda de pacotes em rajadas nas redes IP e redes móveis é um dos fatores mais importantes a serem considerados na análise de sistemas de reconhecimento de voz distribuídos. Perdas de pacotes em rajadas causam uma redução drástica do desempenho do reconhecimento de voz. Neste artigo, nós apresentamos uma técnica nova para a reconstrução dos pacotes perdidos baseada em Redes Neurais e comparamos seu desempenho de reconhecimento com os aqueles obtidos com as técnicas de inserção de zeros e interpolação linear.

Na seção II deste artigo nós fornecemos uma revisão breve dos codecs ITU-T G.723.1 e AMR-NB. Na seção III descrevemos o procedimento de Interpolação das LSFs. Na seção IV tratamos das perdas de pacotes em rajadas nas redes IP e Móveis Celulares. Na seção V, apresentamos a reconstrução de pacotes perdidos usando Inserção de Zeros e Interpolação Linear. Na seção VI, nós propomos uma nova técnica baseada em Redes Neurais a fim reconstruir os pacotes perdidos. As condições experimentais são apresentadas na seção VII. Na Seção VIII, analisamos os resultados de simulação. Finalmente, a seção IX apresenta as conclusões.

II. CODECS ITU-T G.723.1 E AMR-NB

O codec ITU-T G.723.1 permite a codificação de voz a taxas de 6,3 kbit/s ou 5,3 kbit/s [1]. A taxa mais elevada fornece uma voz de melhor qualidade, porém a taxa mais baixa também fornece uma boa qualidade de voz. A diferença entre essas taxas resulta do tipo de excitação a ser utilizada e transmitida para o decodificador. Na taxa de 6,3 kbit/s, o codificador utiliza para a excitação o MP-MLQ (Multi-pulse Maximum Likelihood Quantization), enquanto que na taxa de 5,3 kbit/s é empregado o ACELP (Algebraic Code-Excited Linear Prediction). O codificador opera sobre quadros de 240 amostras cada, o que equivale a 30 ms a uma taxa de amostragem de 8 kHz. Os 10 parâmetros LSF são codificados por um Predictive Split Vector Quantizer em 24 bits/quadro para ambas as taxas de codificação.

O codec de AMR-NB opera-se nas seguintes taxas de bits: 4,75, 5,15, 5,9, 6,7, 7,4, 7,95, 10,2 e 12,2 kbit/s. O AMR-NB é um codificador do tipo ACELP [2]. Opera sobre quadros de voz de 20 ms que correspondem a 160 amostras na freqüência de amostragem de 8 kHz. A análise LP é executada duas vezes por quadro para a taxa do codificador de 12,2 kbit/s e uma vez para as outras taxas. Para a taxa de 12,2 kbit/s, os dois conjuntos de parâmetros LP são convertidos para dois conjuntos de 10 LSFs os quais são conjuntamente quantizados usando-se um Split Matrix Quantization (SMQ) com 38 bits/quadro. Para as outras taxas, o único conjunto de parâmetros LP é convertido para 10 LSFs e quantizado com um Split Vector Quantization. Em 10,2, 7,4, 6,7 e 5,9 as LSFs são quantizas com 26 bits/quadro e em 7,95 kbit/s as LSFs são codificadas com 27 bits/quadro. Nas taxas de 5,15 e 4,75 as LSFs são quantizadas com 23 bits/quadro. Note-se que as diferentes taxas de bits deste codec são geralmente chamadas de modos. A padronização do AMR-NB em 1999 [7] como o codec de voz do GSM representou uma melhoria grande da qualidade da voz para as redes móveis. O codec AMR-NB foi adotado também em 1999 por 3GPP como o codec de voz para o sistema de WCDMA 3G. O codec AMR foi desenvolvido conjuntamente pela Ericsson, Nokia e Siemens [7].

III. INTERPOLAÇÃO DAS LSFS

A Interpolação Linear é uma técnica empregada geralmente em sistemas de reconhecimento de voz distribuídos para interpolar as LSFs decodificadas [5], [8]. Em [6] uma nova técnica foi proposta que supera a Interpolação Linear e por esta razão será usada neste artigo. Esta nova técnica de interpolação é projetada usando um up-sampler seguido por um filtro digital passa-baixa $H(z)$. O up-sampler com fator $r > 1$ (onde r é o fator de interpolação, no caso do ITU-T G.723.1 é 3 e no caso de AMR-NB é 2) insere $r-1$ amostras zeradas equidistantes entre duas amostras consecutivas. O filtro digital passa-baixa $H(z)$ elimina a inserção das imagens (neste caso duas imagens) do espectro original comprimido por um fator r [6].

IV. PERDAS DE PACOTES EM RAJADAS

Embora o IP e as redes móveis sejam completamente diferentes, ambos sofrem de perdas de pacotes em rajadas. Em redes móveis as perdas ocorrem em momentos de forte desvanecimento do sinal, enquanto que em redes IP as perdas de pacotes ocorrerem devido aos congestionamentos. Nós adotamos que exatamente um quadro é encapsulado em um pacote.

Para considerar as características de rajadas do processo de perdas de pacotes, o mesmo foi aproximado por um modelo Markoviano de dois-estados, conhecido também como modelo de Gilbert [9]. Os dois estados referem-se aos eventos “pacote recebido” e “pacote perdido”, respectivamente, p denota a probabilidade da transição do estado “pacote recebido” para o de “pacote perdido”, e q a probabilidade da transição do estado “pacote perdido” para o estado “pacote recebido". A taxa de perda de pacotes (PLR - packet Lost Rate), sabido também como a probabilidade incondicional de perda (ulp - unconditional loss probability) é dado por: $PLR = p/(p+q)$. O comprimento da rajada (*plg* - packet loss gap) conhecido também como o comprimento médio da rajada (B) é dado por $B = 1/(1-clp)$, onde clp (conditional loss probability) é a probabilidade condicional de perda de pacotes, isto é, a probabilidade da transição do estado "pacote perdido" para "pacote perdido" (isto é. $clp = 1-q$). O modelo de perda de pacotes foi simulado neste artigo com as condições de rede usadas em [9] e apresentados na Tabela I.

TABELA I
SIMULAÇÃO DAS CONDIÇÕES DE REDE USANDO O MODELO DE GILBERT.

PLR(%)	clp	B	p	q
0	-	-	0	0
10	0.15	1.18	0.10	0.85
20	0.30	1.43	0.20	0.70
30	0.35	1.54	0.30	0.65
40	0.50	2.00	0.30	0.50

V. Reconstrução usando Inserção de Zeros e Interpolação Linear

Existem algumas aproximações para melhorar o desempenho do sistema de reconhecimento de voz, na presença de imperfeições do canal tais como apagamentos dos quadros. Uma solução simples é inserção dos zeros na posição dos pacotes perdidos. Uma outra aproximação é interpolação linear, entre pacotes recebidos com sucesso (em nosso caso, quadros). O destino recebe, por exemplo, o primeiro conjunto de LSFs quantizadas. Entretanto, devido às imperfeições do canal, não é recebido o segundo conjunto. Na chegada do terceiro conjunto, o receptor pode aproximar o segundo pela interpolação linear do primeiro conjunto com o terceiro. Certamente, a interpolação de mais de um conjunto é praticável em troca de um incremento indesejável de atraso [9]. Para aplicações de redes IP, se n quadros consecutivos de duração t cada um, é perdido, o atraso devido à interpolação é $D_i = nt + RTT/2$, onde RTT (Round-Trip Time) é o tempo para um pacote ir da fonte ao destino e então de volta à fonte. Anote que valores típicos para RTT varia de 10 a 700 ms e de acordo com [9], atrasos aceitáveis para aplicações de VoIP não devem exceder 800 ms [9].

É importante notar que a primeira técnica (Inserção Zero) ignora as características do sinal. Consequentemente, não explora o conhecimento do sinal para melhorar o desempenho do reconhecimento. Por outro lado, o uso da segunda técnica (Interpolação Linear) implica geralmente em longo atraso nos pacotes reconstruídos.

VI. Reconstrução usando Redes Neurais

Pelas razões expostas na seção anterior, nós propusemos neste artigo, uma nova técnica baseada em Redes Neurais para reconstrução dos pacotes perdidos (com a vantagem de usar o conhecimento do comportamento do sinal) e evitar o retardo significativo para a reconstrução do sinal. O atraso da técnica proposta é somente o tempo das Redes Neurais para computar a saída. Este cálculo está baseado nos quadros de LSFs recebidos antes do pacote perdido ou das LSFs interpoladas obtidas antes do pacote perdido que se deseja recuperar.

Na Figura 1 é apresentada a topologia das Redes Neurais escolhida baseado em resultados de simulações obtidas em uma série de estudos preliminares. A camada escondida é composta de 3 neurônios cuja função selecionada para o neurônio foi a tangente hiperbólica. A função linear foi selecionada para o neurônio da camada da saída. Foram utilizadas 10 Redes Neurais com esta topologia, cada uma para uma das 10 LSFs de cada quadro. As 4 entradas de cada Rede Neural são os valores das LSFs em $T-4$, $T-3$, $T-2$ e $T-1$ onde T é o instante em que um quadro é perdido. A saída é a LSF reconstruída em T. Este valor da LSF será usado no sistema de reconhecimento de voz e como uma entrada da rede neural se a LSF de $T+1$ for perdida também. Cada uma das 10 Redes Neurais são treinadas inicialmente com a mesma base de dados usada no treinamento do HMMs (Hidden Markov Models). É interessante observar que quando são recebidos 5 quadros sucessivamente com sucesso, são usados os primeiros 4 pacotes como entradas das Redes Neurais e o quinto pacote como sua saída. Este procedimento tem como única finalidade re-treinar (re-estimar) as Redes Neurais.

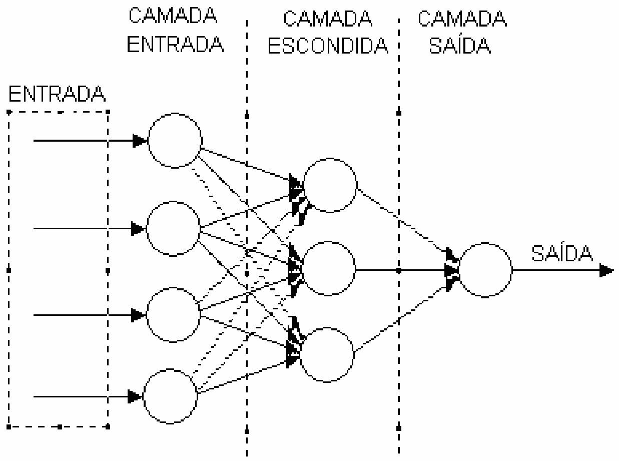


Fig. 1. Topologia das Redes Neurais.

VII. CONDIÇÕES EXPERIMENTAIS

A base de dados usada neste artigo é composta de 50 locutores masculinos e 50 femininos, onde cada locutor fala 1000 sentenças (3.528 palavras) no português Brasileiro. A base de dados foi gravada em um estúdio em uma frequência de amostragem de 16 kHz e em 16 bits por amostra com uma largura de faixa de 50 - 7000 Hertz. Esta base de dados foi filtrada e sub-amostrada para ser compatível com as entradas especificadas pelo ITU-T G.723.1 [1] e pelo AMR-NB [2]. As simulações foram realizadas em um cenário independente do locutor e do texto, e caracterizam o cenário que melhor se aproxima do uso prático do sistema de reconhecimento de voz distribuído. Foi utilizada uma distribuição de 56.25% da base de dados para o treinamento (75 locutores, cada um falando 750 sentenças), 6.25% para testar (25 locutores diferentes, cada um falando 250 sentenças diferentes) e 37.5% da base de dados não foi utilizada. Para garantir a confiança estatística dos resultados, foi empregada a validação cruzada em todas as simulações. O desempenho foi medido nos termos das taxas médias de reconhecimento de palavra ($\overline{WRR}$), desvio padrão (σ) e intervalos de confiança de 95%.

Os extratores de atributos geram um conjunto de 10 atributos mais suas primeiras e segundas derivadas, representando um total de 30 atributos de reconhecimento. Note que os 10 atributos correspondem aos 10 MPCEP convertidos das LSFs quantizadas pelos dois codecs em taxas diferentes. A diferença entre as taxas do ITU-T G.723.1 e AMR-NB afeta significativamente o desempenho obtido com os atributos de reconhecimento, o que será observado nos resultados de simulação. O modelo acústico usa HMMs contínuas de três estados (Hidden Markov Models) com uma mistura de vinte Gaussianas por estado para modelar o fone. Considerando o silêncio estacionário, foi usado um estado com o mesmo número de Gaussians para representá-lo. Os trifones Inter- e Intra-palavra são usados como unidades acústicas. O modelo de linguagem Trigrama para o português Brasileiro foi treinado com um léxico de 60.080 palavras com perplexidade de 307 obtidas de 240.000 sentenças extraídas de um corpus grande de textos do Ceten-Folha [10].

VIII. ANÁLISE DOS RESULTADOS DE SIMULAÇÃO

Os resultados do desempenho são apresentados em cinco tabelas, onde em cada tabela são mostrados o desempenho de reconhecimento para o MPCEP obtido das LSF em diversas taxas dos codificadores ITU-T G.723.1 e AMR-NB para diferentes condições de rede. A tabela II mostra os resultados do reconhecimento para uma rede ideal sem perda dos pacotes. As tabelas III, IV, V e VI mostram os desempenhos de reconhecimento para redes reais com taxas da perda de pacotes PLR e comprimento médio das rajadas B dados por $PLR = 0,\ 10,\ 20,\ 30$ e 40% e $B = 0,\ 1{,}18,\ 1{,}43,\ 1{,}54$ e $2{,}00$, respectivamente. Deve-se observar que em cada caso de teste, os parâmetros do modelo são treinados com o mesmo tipo de atributos (o mesmo tipo de reconstrução), isto é, treinamento e teste estão casados neste sentido. É também importante lembrar que o AMR-NB quando operando em 12,2 kbit/s gera LSFs em 100 Hz, o que evita a necessidade de interpolação das LSFs para esta taxa do codec (para as outras taxas do codec AMR-NB, as LSFs são geradas em 50 Hz e necessitam ser interpoladas para atingir os 100 Hz).

TABELA II
DESEMPENHO DE RECONHECIMENTO PARA REDES SEM PERDAS DE PACOTES.

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	71,32%	1,57%	[70,63% ; 72,01%]
MPCEP - AMR-NB (12,2 kbit/s)	74,10%	1,53%	[73,42% ; 74,78%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	72,87%	1,58%	[72,18% ; 73,56%]
MPCEP - AMR-NB (7,95 kbit/s)	72,89%	1,58%	[72,20% ; 73,58%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	72,53%	1,61%	[71,82% ; 73,24%]

TABELA III
DESEMPENHO DE RECONHECIMENTO PARA REDES COM PLR=10% E B=1,18.

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
Inserção de Zeros			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	66,21%	1,59%	[65,51% ; 66,91%]
MPCEP - AMR-NB (12,2 kbit/s)	70,01%	1,54%	[69,34% ; 70,69%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	68,12%	1,60%	[67,42% ; 68,82%]
MPCEP - AMR-NB (7,95 kbit/s)	68,15%	1,59%	[67,45% ; 68,85%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	67,09%	1,64%	[66,37% ; 67,81%]
Interpolação Linear			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	67,52%	1,59%	[66,82% ; 68,22%]
MPCEP - AMR-NB (12,2 kbit/s)	71,45%	1,54%	[70,78% ; 72,13%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	69,53%	1,59%	[68,83% ; 70,23%]
MPCEP - AMR-NB (7,95 kbit/s)	69,55%	1,59%	[68,85% ; 70,25%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	68,51%	1,63%	[67,80% ; 69,23%]
Redes Neurais			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	67,54%	1,58%	[66,85% ; 68,23%]
MPCEP - AMR-NB (12,2 kbit/s)	71,49%	1,54%	[70,82% ; 72,17%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	69,54%	1,59%	[68,84% ; 70,24%]
MPCEP - AMR-NB (7,95 kbit/s)	69,58%	1,59%	[68,88% ; 70,28%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	68,52%	1,62%	[67,81% ; 69,23%]

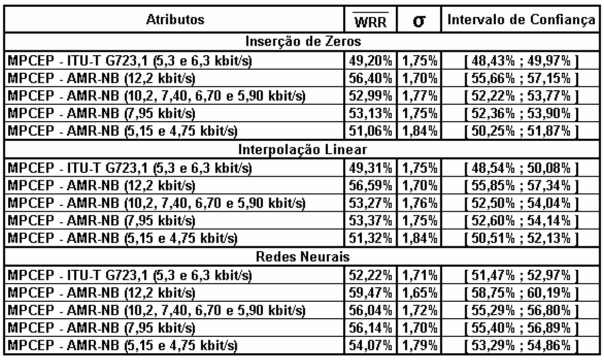
TABELA IV
DESEMPENHO DE RECONHECIMENTO PARA REDES COM PLR=20% E B=1,43.

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
Inserção de Zeros			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	61,57%	1,64%	[60,85% ; 62,29%]
MPCEP - AMR-NB (12,2 kbit/s)	65,82%	1,58%	[65,13% ; 66,51%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	63,71%	1,65%	[62,99% ; 64,43%]
MPCEP - AMR-NB (7,95 kbit/s)	63,77%	1,64%	[63,05% ; 64,49%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	62,22%	1,70%	[61,48% ; 62,97%]
Interpolação Linear			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	62,63%	1,63%	[61,92% ; 63,35%]
MPCEP - AMR-NB (12,2 kbit/s)	66,84%	1,58%	[66,15% ; 67,53%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	64,72%	1,65%	[64,00% ; 65,44%]
MPCEP - AMR-NB (7,95 kbit/s)	64,78%	1,63%	[64,07% ; 65,50%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	63,21%	1,70%	[62,47% ; 63,96%]
Redes Neurais			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	63,12%	1,61%	[62,42% ; 63,83%]
MPCEP - AMR-NB (12,2 kbit/s)	67,37%	1,56%	[66,69% ; 68,06%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	65,28%	1,64%	[64,56% ; 66,00%]
MPCEP - AMR-NB (7,95 kbit/s)	65,34%	1,61%	[64,64% ; 66,05%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	63,71%	1,69%	[62,97% ; 64,45%]

TABELA V
DESEMPENHO DE RECONHECIMENTO PARA REDES COM PLR=30% E B=1,54.

Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
Inserção de Zeros			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	57,79%	1,68%	[57,06% ; 58,53%]
MPCEP - AMR-NB (12,2 kbit/s)	62,01%	1,63%	[61,30% ; 62,73%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	59,64%	1,69%	[58,90% ; 60,38%]
MPCEP - AMR-NB (7,95 kbit/s)	59,72%	1,68%	[58,99% ; 60,46%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	58,10%	1,75%	[57,33% ; 58,87%]
Interpolação Linear			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	58,57%	1,68%	[57,84% ; 59,31%]
MPCEP - AMR-NB (12,2 kbit/s)	62,81%	1,63%	[62,10% ; 63,53%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	60,27%	1,69%	[59,53% ; 61,01%]
MPCEP - AMR-NB (7,95 kbit/s)	60,37%	1,68%	[59,64% ; 61,11%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	58,68%	1,75%	[57,91% ; 59,45%]
Redes Neurais			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	59,93%	1,66%	[59,20% ; 60,66%]
MPCEP - AMR-NB (12,2 kbit/s)	64,49%	1,60%	[63,79% ; 65,19%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	63,91%	1,67%	[63,18% ; 64,64%]
MPCEP - AMR-NB (7,95 kbit/s)	63,99%	1,66%	[63,26% ; 64,72%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	62,41%	1,73%	[61,65% ; 63,17%]

TABELA VI
DESEMPENHO DE RECONHECIMENTO PARA REDES COM PLR=40% E B=2,00.



Atributos	$\overline{WRR}$	σ	Intervalo de Confiança
Inserção de Zeros			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	49,20%	1,75%	[48,43% ; 49,97%]
MPCEP - AMR-NB (12,2 kbit/s)	56,40%	1,70%	[55,66% ; 57,15%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	52,99%	1,77%	[52,22% ; 53,77%]
MPCEP - AMR-NB (7,95 kbit/s)	53,13%	1,75%	[52,36% ; 53,90%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	51,06%	1,84%	[50,25% ; 51,87%]
Interpolação Linear			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	49,31%	1,75%	[48,54% ; 50,08%]
MPCEP - AMR-NB (12,2 kbit/s)	56,59%	1,70%	[55,85% ; 57,34%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	53,27%	1,76%	[52,50% ; 54,04%]
MPCEP - AMR-NB (7,95 kbit/s)	53,37%	1,75%	[52,60% ; 54,14%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	51,32%	1,84%	[50,51% ; 52,13%]
Redes Neurais			
MPCEP - ITU-T G723,1 (5,3 e 6,3 kbit/s)	52,22%	1,71%	[51,47% ; 52,97%]
MPCEP - AMR-NB (12,2 kbit/s)	59,47%	1,65%	[58,75% ; 60,19%]
MPCEP - AMR-NB (10,2, 7,40, 6,70 e 5,90 kbit/s)	56,04%	1,72%	[55,29% ; 56,80%]
MPCEP - AMR-NB (7,95 kbit/s)	56,14%	1,70%	[55,40% ; 56,89%]
MPCEP - AMR-NB (5,15 e 4,75 kbit/s)	54,07%	1,79%	[53,29% ; 54,86%]

Dos resultados da simulação fica claro que a Inserção de Zeros é definitivamente a pior aproximação para a solução da perda de pacotes. Agora comparando a Inserção de Zeros, a Interpolação Linear e a Redes Neurais para a reconstrução de pacotes perdidos de LSFs nas tabelas III, IV, V e VI, pode-se ver que a técnica proposta que usa Redes Neurais supera as duas outras técnicas em todos os casos. Entretanto, as melhorias são somente significativas nas tabelas V e VI, correspondendo à perda de pacotes - PLR - de 30% e 40%, respectivamente, onde as condições das redes IP e das redes móveis são mais severas. No caso onde $PLR = 40\%$ (Tabela VI), o esquema novo fornece ganhos de reconhecimento de aproximadamente 3% quando comparado com a técnica da Interpolação Linear.

Agora comparando o ITU-T G.723.1 nas taxas 6,3 e 5,3 kbit/s com o codec de AMR-NB nas taxas similares (6,7 e 5,9 kbit/s), em toda as condições de rede, o AMR-NB fornece um ganho em torno de 1,50%. É também muito significativo notar que os intervalos de confiança de 95% não se sobrepõem. Além disso, é importante observar que o AMR-NB em suas taxas mais baixas (5,15 e 4,75 kbit/s) supera o codec de ITU-T G.723.1 que opera em 6,3 e 5,3 kbit/s para todos os valores de PLR. Outra vez, seus intervalos de confiança de 95% não se sobrepõem. Note que as LSFs de onde os atributos de reconhecimento são extraídos, são codificados em uma taxa de bits mais elevada pelo AMR-NB em comparação ao ITU-T G.723.1. Finalmente, está claro que as Redes Neurais são uma técnica atrativa para a reconstrução de pacotes perdidos para ambos os codificadores de voz.

IX. CONCLUSÕES

Neste artigo, nós realizamos diversas experiências importantes em Reconhecimento de Voz Contínuo Distribuído com amplo vocabulário no português Brasileiro. Nós propusemos o uso de Redes Neurais para a reconstrução de pacotes perdidos em sistemas Móveis e redes IP. Comparando com a Inserção de Zeros e a técnica de Interpolação Linear, as Redes Neurais mostraram ser o melhor método para reconstruir pacotes perdidos em sistemas de Reconhecimento de Voz Distribuído que empreguem os codecs ITU-T G.723.1 ou AMR-NB, especialmente em condições severas de perda de pacotes. Além disso, nós mostramos que o AMR-NB que opera em uma taxa de bits mais baixa supera o codec ITU-T G.723.1 nas taxas de reconhecimento, sem sobreposição dos seus intervalos de confiança em 95%, em todas as condições da rede.

REFERÊNCIAS

[1] ITU-T Recommendation G.723.1, “Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s,” Março 1996.

[2] 3GPP TS 26.071 V6.0.0, “Mandatory speech CODEC speech processing functions, AMR speech CODEC - General description,” Dezembro, 2004.

[3] 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.

[4] V. F. S. Alencar and A. Alcaim, “Transformations of LPC and LSF Parameters to Speech Recognition Features”, Proceedings of the ICAPR, Bath, UK, Agosto 2005.

[5] V. F. S. Alencar and A. Alcaim, “Features Interpolation Domain for Distributed Speech Recognition and Performance for ITU-T G.723.1 CODEC”, Proceedings of the ICSLP, Antwerp, BE, Agosto 2007.

[6] V. F. S. Alencar and A. Alcaim, “Digital Filter Interpolation of Decoded LSFs for Distributed Continuous Speech Recognition”, Electronics Letters, vol.44, issue:17, pp.1039-1040, Agosto 2008.

[7] K. Järvinen, “Standardisation of the Adaptive Multi-rate Codec,” European Signal Processing Conference (EUSIPCO), Tampere, Finland, 4–8 Setembro 2000.

[8] H. K. Kim and R. V. Cox, “A Bitstream-Based Front-End for Wireless Speech Recognition on IS-136 Communications System,” IEEE Trans. On Speech and Audio Processing, vol. 9, pp. 558-568, Jullho 2001.

[9] J. Wang and J. Gibson, “Parameter interpolation to enhance the frame erasure robustness of CELP coders in packet networks”, Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 2001.

[10] “Corpus de Extractus de Textos Eletrônicos Nilc/ Folha de São Paulo (Ceten-Folha)”, 14 Novembro 2005.