Referências Bibliográficas

[1] V. Zue, "Conversational interfaces: advances and challenges," *Proc. of EuroSpeech-97*, vol. V, pp. KN 9–18, 1997.

[2] J. Bellegarda, "Statistical techniques for robust asr: review and perspectives," in *Fifth European Conference on Speech Communication and Technology*, pp. KN 33–36, 1997.

[3] M. Rosenzweig and A. Leiman, *Psicología fisiológica*. McGraw-Hill Madrid, second edition ed., 1992.

[4] D. Halliday, R. Resnick, and M. Krane, *Física para estudiantes deficiencias e ingeniería*, vol. 1. Parte, 1966.

[5] L. Rocha, *Procesamiento de la voz*. Kapelusz, 1987.

[6] A. Alcaim and C. Oliveira, *fundamentos do processamento de sinais de voz e imagem*. Interciencia e PUC Rio, 2012.

[7] O. Dekel, J. Keshet, and Y. Singer, "An online algorithm for hierarchical phoneme classification," *Machine Learning for Multimodal Interaction*, pp. 146–158, 2005.

[8] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. Prentice hall, 1993.

[9] R. Cole, J. Mariani, H. Uszkoreit, G. Varile, A. Zaenen, and A. Zampolli, *Survey of the state of the art in human language technology*, vol. 12. Cambridge University Press, 1998.

[10] J. Mariani, "Recent advances in speech processing," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pp. 429–440, IEEE, 1989.

[11] J. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215–1247, 1993.

[12] P. Moreno, *Speech recognition in noisy environments*. PhD thesis, Carnegie Mellon University, 1996.

[13] F. Sepúlveda Sepúlveda *et al.*, *Extracción de parámetros de señales de voz usando técnicas de análisis en tiempo-frecuencia*. PhD thesis, Universidad Nacional de Colombia-Sede Manizales, 2004.

[14] S. Mitra and Y. Kuo, *Digital signal processing: a computer-based approach*, vol. 2. McGraw-Hill New York, 2006.

[15] J. Siqueira, "Reconhecimento de voz contínua com atributos mfcc, ssch e pncc, wavelet denoising e redes neurais," Master's thesis, Pontifícia Universidade Católica do Rio de Janeiro, 2011.

[16] A. Oppenheim, R. Schafer, J. Buck, *et al.*, *Discrete-time signal processing*, vol. 2. Prentice hall Englewood Cliffs, NJ:, 1999.

[17] O. Luis, S. Sergio, and B. Ricardo, "Reconocimiento de voz usando segmentación de energía usando modelos ocultos de markov de densidad continua," 2004.

[18] X. Huang, A. Acero, H. Hon, *et al.*, *Spoken language processing*, vol. 15. Prentice Hall PTR New Jersey, 2001.

[19] K. Lee, "Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 38, no. 4, pp. 599–609, 1990.

[20] J. Baker, "The dragon system–an overview," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 23, no. 1, pp. 24–29, 1975.

[21] F. Jelinek, L. Bahl, and R. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech," *Information Theory, IEEE Transactions on*, vol. 21, no. 3, pp. 250–256, 1975.

[22] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[23] M. John, F. Gales, *et al.*, *Model-Based Techniques For Noise Robust Speech Recognition*. PhD thesis, Europe PubMed Central, 1995.

[24] J. Deller, J. Proakis, and J. Hansen, *Discrete-time processing of speech signals*. Macmillan publishing company New York, 1993.

[25] L. Baum, "An equality and associated maximization technique in statistical estimation for probabilistic functions of markov processes," *Inequalities*, vol. 3, pp. 1–8, 1972.

[26] S. C. B. D. Santos, *Reconhecimento de Voz Contínua para o Português Utilizando Modelos de Markov Escondidos*. PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro, 1997.

[27] G. Forney Jr, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[28] C. Martins, A. Teixeira, and J. Neto, "Language models in automatic speech recognition," *Electrónica e Telecomunicações*, vol. 4, no. 4, pp. 428–432, 2011.

[29] S. Patra, "Robust speaker identification system," Master's thesis, Super Computer Education and Reseach Center, 2007.

[30] M. Zanuy, *Tratamiento Digital de Voz e Imagen y aplicación a la multimedia*. Marcombo, 2000.

[31] G. Blanchet and M. Charbit, *Digital signal and image processing using Matlab*, vol. 5. Wiley-ISTE, 2010.

[32] H. Moore, V. Olguín, and R. Nuño, *Matlab para ingenieros*. Pearson Prentice Hall, 2007.

[33] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*, vol. 3. Cambridge University Engineering Department, 2002.

[34] W. Han, C. Chan, C. Choy, and K. Pun, "An efficient mfcc extraction method in speech recognition," in *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, pp. 4–pp, IEEE, 2006.

[35] J. Deller, J. H. Hansen, and P. J.G, "The relation of pitch to frequency," *IEEE Press*, vol. 1, p. 936, 2000.

[36] R. T. Tevah, "Iimplementação de um sistema de reconhecimento de fala contínua com amplo vocabulário para o português brasileiro," Master's thesis, Universidade Federal do Rio de Janeiro, 2006.

[37] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in

ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW), 2000.

[38] C. Kim and R. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," *INTERSPEECH-2009*, vol. 1, pp. 28–31, 2009.

[39] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," *Auditory physiology and perception*, vol. 83, pp. 429–446, 1992.

[40] D. Wang and G. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. IEEE Press, 2006.

[41] M. Slaney *et al.*, "An efficient implementation of the patterson-holdsworth auditory filter bank," *Apple Computer, Perception Group, Tech. Rep*, vol. 35, p. 8, 1993.

[42] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, "Implementing a gammatone filter bank," *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, vol. 1, pp. 1–5, 1988.

[43] G. Sárosi, M. Mozsáry, P. Mihajlik, and T. Fegyó, "Comparison of feature extraction methods for speech recognition in noise-free and in traffic noise environment," in *Speech Technology and Human-Computer Dialogue (SpeD), 2011 6th Conference on*, pp. 1–8, IEEE, 2011.

[44] C. Kim and R. Stern, "Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4574–4577, IEEE, 2010.

[45] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition," in *Proc. ICSLP*, vol. 3, pp. 869–872, 2000.

[46] A. Gallardo Antolín, *Reconocimiento de habla robusto frente a condiciones de ruido aditivo y convolutivo*. PhD thesis, Universidad Politécnica de Madrid, 2002.

[47] F. Liu, *Environmental adaptation for robust speech recognition*. PhD thesis, Citeseer, 1994.

[48] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[49] D. A. de Oliveira Santos, "Reconhecimento de voz em presença de ruído," Master's thesis, Pontifícia Universidade Católica do Rio de Janeiro, 2001.

[50] S. Vaseghi, *Advanced digital signal processing and noise reduction*. Wiley, 2008.

[51] J. Segura, C. Benitez, A. De La Torre, and A. Rubio, "Feature extraction combining spectral noise reduction and cepstral histogram equalization for robust asr," in *Proc. ICSLP 02*, pp. 225–228, 2002.

[52] O. Borras Gene, "Reductor de ruido mediante resta espectral en entorno matlab," in *EU IT. Telecomunicacion (UPM)*, 2006.

[53] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[54] Y. Salimpour and M. Abolhassani, "Auditory wavelet transform based on auditory wavelet families," in *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*, pp. 1731–1734, IEEE, 2006.

[55] A. Gallardo Antolín, *Reconocimiento de habla robusto frente a condiciones de ruido aditivo y convolutivo*. PhD thesis, Universidade Politécnica de Macrid, 2002.

[56] C. Medina, A. Alcaim, and J. Apolinario Jr, "Wavelet denoising of speech using neural networks for," *Electronics letters*, vol. 39, no. 25, pp. 1869–1871, 2003.

[57] D. Donoho and I. Johnstone, "Threshold selection for wavelet shrinkage of noisy data," in *Engineering in Medicine and Biology Society, 1994. Engineering Advances: New Opportunities for Biomedical Engineers. Proceedings of the 16th Annual International Conference of the IEEE*, pp. A24–A25, IEEE, 1994.

[58] O. Farooq and S. Datta, "Wavelet-based denoising for robust feature extraction for speech recognition," *Electronics Letters*, vol. 39, no. 1, pp. 163–165, 2003.

[59] J. K. Siquiera, "Reconhecimento de voz contínua com atributos mfcc, ssch e pncc, wavelet denoisign e redes neurais," Master's thesis, Pontífica Universidade Católica do Rio de Janiero, 2011.

[60] *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) Training and Test Data*, 1993.

[61] A. Varga, H. Steeneken, M. Tomlinson, and D. Jones, "The noisex-92 study on the effect of additive noise on automatic speech recognition," *DRA Speech Research Unit, Malvern, England, Tech. Rep*, 1992.

[62] M. Harvilla and R. Stern, "Histogram-based subband powerwarping and spectral averaging for robust speech recognition under matched and multistyle training," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4697–4700, IEEE, 2012.

[63] A. De La Torre, A. Peinado, J. Segura, J. Pérez-Córdoba, M. Benítez, and A. Rubio, "Histogram equalization of speech representation for robust speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 355–366, 2005.

[64] R. Balchandran and R. Mammone, "Non-parametric estimation and correction of non-linear distortion in speech systems," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2, pp. 749–752, IEEE, 1998.

[65] B. C. D. L. T. A. R. A. Segura, J.C. and J. Ramirez, "Efectos no lineales del entorno acústico en parametrizaciones para reconocimiento automático de voz basadas en mfcc," in *Tercera Jotnada en tecnología del habla - Valencia*, 2004.

[66] G. Saon, S. Dharanipragada, and D. Povey, "Feature space gaussianization," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1, pp. I–329, IEEE, 2004.

[67] S. Haikin, *Neural networks: a comprehensive foundation*, vol. 2. Pearson Education, 1998.

[68] D. Matich, "Redes neuronales: Conceptos básicos y aplicaciones," in *Cátedra de Informática Aplicada a la Ingeniería de Procesos–Orientación I*, 2001.

[69] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural networks*, vol. 4, no. 2, pp. 251–257, 1991.

A
Algoritmo de Baum-Welch

Esta técnica utiliza os algoritmos precisos para fazer o calculo ou estimação dos parâmetros que definem o modelo. Trara-se basicamente de um processo iterativo que maximiza em forma local a probabilidade de que uma sequência de observação seja gerada por um modelo particular $[P(O|\lambda)]$ e que garante em certa forma a convergência do processo a partir de um modelo inicial aleatório.

Um dos métodos mais conhecidos para realizar esta tarefa é a técnica Expectation-Maximization(EM) e como uma especialização da mesma, o algoritmo de Baum-Welch. Ele é apropriado justamente para estimar parâmetros quando não se conhece uma das variáveis (neste caso, a sequência de estados s). Apesar do resultado não ser necessariamente o máximo global, ele garante a convergência para um máximo local, definindo os parâmetros como apresenta-se a seguir:

- Probabilidade de transição de estado (matriz A)

$$a_{ij} = \frac{\textit{Número esperado de transições de } i \textit{ para } j}{\textit{Número esperado de transições desde } i}$$

- Probabilidade de emissão (Matriz B)

$$b_{j(k)} = \frac{\textit{Número esperado de vezes no estado } j \textit{ com o simbolo } v_k \textit{ observado}}{\textit{Número de vezes no estado } j}$$

- Vetor de probabilidade inicial

$$\prod_i = \textit{probabilidade de iniciar no estado } i$$

Assim, dada uma sequência, procura-se mediante o método EM obter o modelo HMM que tenha a probabilidade de gerar a sequência indicada, utilizando as formulas apresentadas em [59] (vide apêndice E) e o algoritmo geral apresentado a seguir

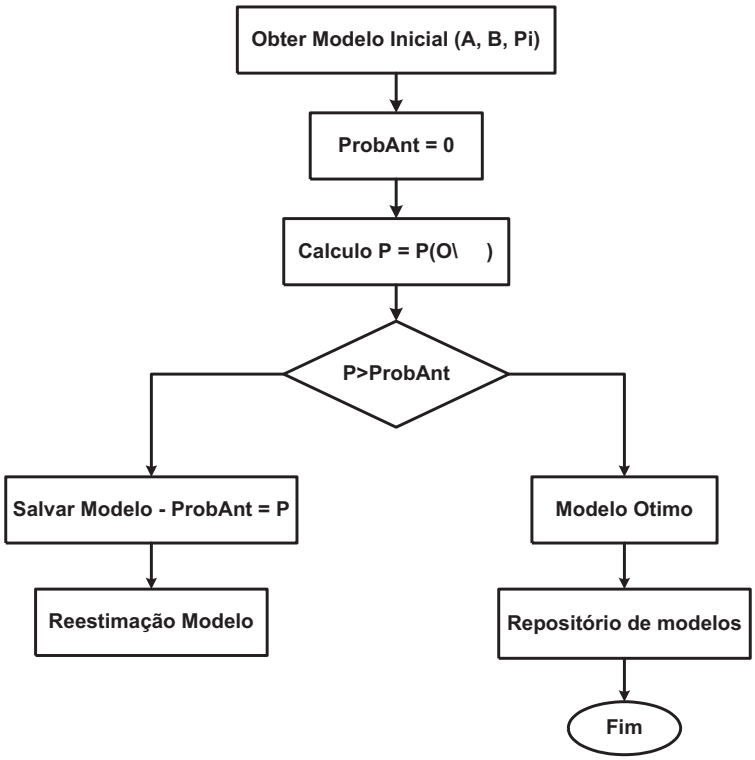

Figura A.1: Diagrama de fluxo do método EM.

Por conseguinte, o algoritmo EM propõe obter em forma iterativa modelos λ_n para uma sequência de observações mostra, e ir comparando assim as probabilidades de geração dos mesmos, que estão crescendo até atingir o máximo local para essa probabilidade. Atingindo este máximo (passo garantido pela convergência do método) toma-se ao modelo final como o que mais provavelmente possa gerar a sequência de observações pela mostra.

B
Algoritmo de Viterbi

Define-se $v_t(j)$ como a probabilidade máxima de que um modelo de Markov HMM, M atinga o estado s no instante t, emitindo a observação O

$$v_t(j) = \max_{s_1,s_2...s_t} P(O_1, O_2, ..., s_1, s_2, ..., s_t) \tag{B-1}$$

$v_t(j)$ pode-se calcular como

$$v_t(j) = \max_{s_1,...s_{t-1}} P(O_1, O_{t-1}, ..., s_1, s_2, ..., s_{t-1}).A_{s_{t-1}s_t}B_{s_tO_t} \tag{B-2}$$

$$v_t(j) = \max_{s_{t-1}\in S} \max_{s_1,s_2...s_{t-1}} P(O_1, O_{t-1}, ..., s_1, s_2, ..., s_{t-1}).A_{s_{t-1}s_t}B_{s_tO_t} \tag{B-3}$$

$$= \max_{s_{t-1}\in S} V(s_{t-1}, t-1).A_{s_{t-1}s_t}B_{s_tO_t} \tag{B-4}$$

A equação B-5 traduz a ideia geral do algoritmo: considerar apenas o máximo dos caminhos passados.

$$v_t(j) = \begin{cases} \Pi_s B_s, O_1 & \text{si} \quad t = 1 \\ \max_{s_{t-1}\in S} V(s_{t-1}, t-1).A_{s_{t-1}s_t}B_{s_tO_t} & \text{si} \quad t > 1 \end{cases} \tag{B-5}$$

Para que a indução se feche, resta apenas o termo inicial, obtido facilmente com as propriedades do HMM.

$$v_1(j) = P(O_1, s_1 = j) = P(O_1 \mid s_1 = j)P(s_1 = j) = b_j(O_1)\Pi_j \tag{B-6}$$

Assim a partir de cada $v_{t-1}(i)$, é possível acumular os nós anteriores de maior probabilidade na variável $B_t(i)$ através das equações 2-21 até 2-24.

C
Matlab

Matlab é o nome abreviado de "MATrix LABoratory"é um programa pago, fechado e multiplataforma que utiliza uma linguagem de alto nível e um ambiente iterativo baseado em matrizes para cálculos científicos e de engenharia, que permite fazer tarefas intensas com maior velocidade do que as linguagens de programação geralmente utilizadas.

Matlab especializa se em cálculos numéricos com vetores e matrizes, com a possibilidade de manipular grandes quantidades de dados com poucos comandos, permitindo trabalhar também com outras estruturas de informação, mesmo que cada objeto dessa estrutura seja considerado como um arreglo.

Outra das possibilidades de Matlab é a realização de uma grande variedade de grafos em dois e três dimensões

Sua linguagem de programação é uma ferramenta de alto nível para desenvolver aplicações técnicas através de um código chamado M-code e varias bibliotecas especializadas chamadas toolboxes que dão suporte a números complexos e uma serie de funções matemáticas, fornecidas para o processamento de sinais além de funções para o processamento digital.

A figura C.1 apresenta a interface gráfica do software.

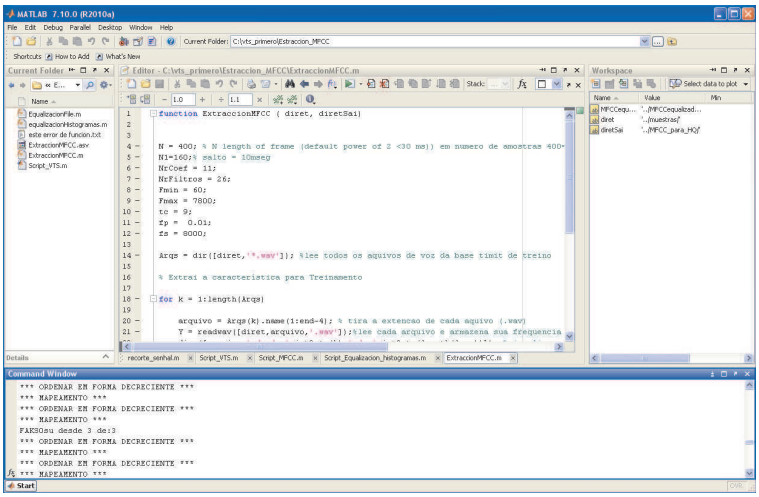

Figura C.1: Interface gráfica do programa Matlab.

D
HTK

Esta ferramenta computacional é definida como um conjunto portátil de ferramentas de software gratuito, aberto e multiplataforma para construção e manipulação de HMMs. Criado pelo professor Steve Young em 1989 a fim de servir como mecanismo de pesquisa em reconhecimento de voz, agora pode-se utilizar em qualquer área de conhecimento como reconhecimento de caracteres, sequenciamento de DNA, análises de vibrações, entre outros, estabelecendo a restrição de que o problema a resolver possa ser focado como um processo estocástico Markoviano.

O HTK é controlado por modulo de librarias com códigos fonte disponíveis em C, as ferramentas contidas nele proveem sofisticadas funcionalidades para análise de fala, treino de HMMs, teste e análise de resultados. Para utilizar qualquer destas ferramentas deve-se ter em conta os seguintes aspectos:

– Linha de comando com a interface do sistema operativo, através da qual são chamados seus programas, por exemplo

  HCopy -C configuração.txt -S arquivo.txt

  Neste exemplo extraem-se os atributos MFCC dos arquivos listados em *arquivo.txt* usando uns parâmetros previamente estabelecidos no arquivo.

  configuração.txt O resultado tem que possuir extensão "mfcc"mostrando assim que os dados correspondem a vetores codificados segundo coeficientes cepstrais de frequência MEL

– Modulo de librarias onde tem-se distintos tipos de arquivos, cada uma com um conjunto de instruções a fim de conseguir uma função específica nas ferramentas disponíveis, por exemplo

  Hlabel a qual controla os arquivos de etiquetas

  Na tabela. D.1 resume-se as ferramentas mais utilizadas no caso de reconhecimento de voz, dividendo-as em quatro grupos.

Devido à grande quantidade de arquivos que tem-se que manipular, os processos podem ser longos e complexos na sua construção. Portanto os

GRUPO	FERRAMENTA
Ferramenta de preparação de dados	HSlab, HCopy, HLed
Ferramenta de treinamento	HInit, HRest, HCompv HHed
Ferramenta de reconhecimento	HVite, HParse, HSgen
Ferramenta de análise	HResult

Tabela D.1: Grupo de ferramentas de HTK utilizadas nas aplicações de reconhecimento de voz.

programadores sugeriram o desenvolvimento de uma interface amigável, a fim de guiar ao usuário passo a passo nas diferentes fases empregadas para a construção dos reconhecedores propostos.

Esta interface Fig. D.1 foi feita em código de Visual Basic, conseguindo que a informação fornecida pelo usuário seja limitada ao vocabulário do reconhecedor, com características típicas de cada aplicação.

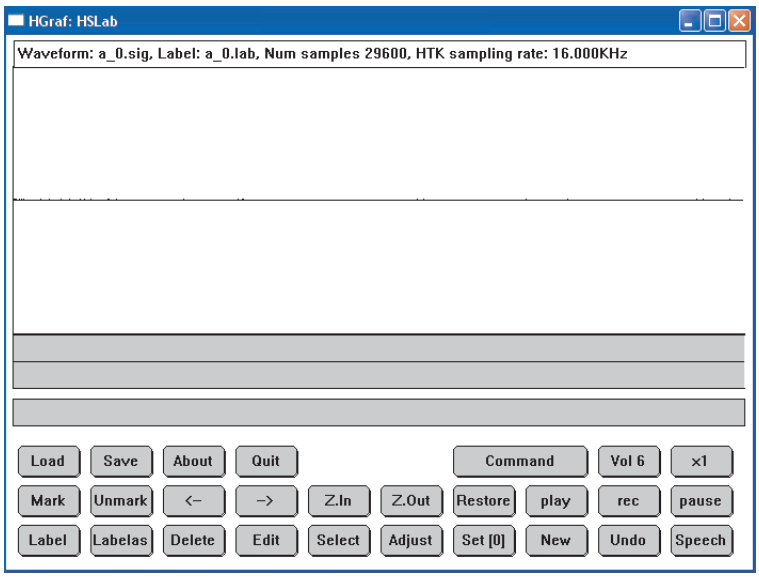

Figura D.1: Interface gráfica ao invocar a ferramenta Hslab.

A versão disponível do HTK é a 3.3, liberada em 25/07/2005. Contém extensiva documentação e exemplos em um documento conhecido como HTK Book.

E
Ruído branco gaussiano

O ruído branco aditivo gaussiano (AGWN), por suas siglas em inglês, é utilizado frequentemente nas experiencias de reconhecimento de voz a fim de contaminar de forma controlada o sinal de voz por varias razões:

- Esse tipo de ruído afeta toda a faixa de frequência, devido ao fato de que é um processo estocástico no sentido amplio (ESA), com media nula e densidade espectral de potencia (DEP) constante ao longo de todo o espectro do sinal de voz pelo qual é um dos ruídos que afeta de forma mais severa a voz.
- O ruído AGWN constitui um modelo aceitável para muitas fontes de ruído de banda larga encontradas na pratica, um exemplo é o ruído térmico dos dispositivos usados nos equipamentos.
- O AGWN utiliza-se frequentemente para avaliar os sistema de reconhecimento de voz, pelo qual pode ser considerado uma fonte de ruído estandar para os testes de reconhecimento.

E.1
Confecção do Sinal Ruidoso

Os sinais ruidosos utilizados neste trabalho são formados a partir da adição do ruído ao sinal limpo nas diferentes RSRs desejadas. Para tanto, antes de efetuá-la, deve-se multiplicar o sinal de ruído por um determinado fator, da seguinte forma

$$r(t) = \beta \cdot ra(t) \tag{E-1}$$

onde $ra(t)$ é o sinal de ruído original (antes da multipllicação pelo fator adaptativo), β é Fator multiplicativo e $r(t)$ é o sinal de ruído que atende à RSR desejada.

Sabe-se que

$$RSR = 10 \log \left(\frac{E_s}{E_r} \right) \tag{E-2}$$

onde E_r é a energia média do ruído desejado. E_s Energia média do sinal limpo.

Por tanto

$$\frac{RSR}{10} = \log \left(\frac{E_s}{E_r} \right) \tag{E-3}$$

Assim, isolando a energia média do ruído tem se

$$E_r = \frac{E_s}{10^{\frac{RSR}{10}}} \tag{E-4}$$

Sabe se que a energia média de qualquer sinal $x(t)$ é dada por

$$E = \frac{1}{N} \sum_{t=1}^{N} \mid x(t) \mid^2 \tag{E-5}$$

de acordo com isso aplicando E-5 em E-1 tem-se

$$E_{r(t)} = \frac{1}{N} \sum_{t=1}^{N} \mid \beta \cdot ra(t) \mid^2 \tag{E-6}$$

igualando E-4 a E-6 e sabendo que β é independente de t

$$\beta - \frac{1}{\sqrt{E_{ra(t)}}} \sqrt{\frac{E_s}{10^{\frac{RSR}{10}}}} \tag{E-7}$$

onde $E_{ra(t)}$ e a energia média do sinal de ruído original, tendo assim que o ruido desejado é dado por

$$r(t) = \left(\frac{1}{\sqrt{E_{ra(t)}}} \sqrt{\frac{E_s}{10^{\frac{RSR}{10}}}} \right) ra(t) \tag{E-8}$$