SpeakerID – Toolkit para Processamento e Modelagem de Características de Alto Nível para Reconhecimento Automático de Locutor

Cristian Keil de Abreu, André Gustavo Adami

Departamento de Informática – Universidade de Caxias do Sul (UCS)
Caixa Postal 1352 – 95001-970 – Caxias do Sul – RS – Brazil

{ckabreu,agadami}@ucs.br

Abstract. *This paper describes a toolkit for feature extraction and modeling of high-level features for speaker recognition systems. In addition, the toolkit provides tools to evaluate features and models according to the NIST evaluation paradigm (commonly used as reference for evaluating speaker recognition systems). The toolkit was implemented in Perl and C languages and uses several open-source software for process scheduling and feature extraction. Some results of the toolkit on the 2001 NIST competition are presented in this paper.*

Resumo. *Este artigo descreve um toolkit para extração e modelagem de características de alto nível para sistemas de reconhecimento do locutor. Além disso, o toolkit oferece ferramentas para avaliar as características e modelos conforme o paradigma de avaliação do NIST (comumente utilizado como referência de avaliação de tais sistemas). O toolkit foi implementado nas linguagens Perl e C e utiliza diversos softwares livres para tarefas de escalonamento de processos e extração de características. Alguns resultados do toolkit na competição do NIST realizada em 2001 são apresentados neste artigo.*

1. Introdução

O ser humano possui uma capacidade de identificar indivíduos pelas características da fala [Schmidt-Nielsen and Crystal, 1998], mesmo prejudicada pelos ruídos do ambiente ou diferente contexto, através do uso de diferentes níveis de informações [Lavner et al., 2000]. Dependendo da pessoa, o sotaque ou tipo de risada podem ser utilizados para fornecer informação extra sobre a identidade de quem está falando (o qual será tratado como "locutor" no restante do artigo). Inspirados por esta capacidade humana em reconhecer a identidade do locutor e pelas diversas possibilidades de sua aplicação, sistemas de Reconhecimento Automático de Locutor (RAL) estão sendo desenvolvidos por pesquisadores [Furui, 2005]. Entretanto, apesar da grande quantidade de informações existente na voz, a maioria destes sistemas utiliza apenas uma única fonte de informação para realizar o reconhecimento: características físicas do sinal de voz [Bimbot et al., 2004]. Por isso, o acréscimo de novas fontes de informações existentes no sinal de fala aumentará o desempenho dos sistemas de RAL, como mostrado em diversos trabalhos [Adami, 2004; Andrews et al., 2001; Farahani et al., 2004].

Uma das fontes de informação que tem sido reconhecida como discriminante de locutores é a prosódia. Motivado pelo poder discriminatório da prosódia (devido às diferenças na produção de entonação, ênfase, ritmo), diversos trabalhos foram apresentados nos últimos anos [Adami, 2004; Farahani et al., 2004; Hannani and Petrovska-Delacrétaz, 2005]. Tais trabalhos são caracterizados pela utilização de

medidas físicas da freqüência fundamental (medida correlata do tom, representada por F0) e energia (medida correlata da intensidade) do sinal de voz, e uso de modelos probabilísticos e lingüísticos para construção de modelos do locutor. Por isso, o objetivo deste trabalho é de apresentar um novo toolkit, chamado SpeakerID, que permite a geração e modelagem de características de alto-nível (especialmente, as derivadas de prosódia). Para avaliar os resultados obtidos, o paradigma de avaliação proposto pelo *National Institute of Standards and Technology* (NIST) para o reconhecimento de locutor de 2001 [Martin, 2001] está sendo utilizado (devido a sua aceitação internacional como medida de desempenho para este tipo de tecnologia).

Este artigo está organizado da seguinte maneira: Seção 2 descreve as características do toolkit. Na Seção 3, alguns resultados preliminares do toolkit são apresentados. Considerações finais são descritas na Seção 4.

2. Toolkit SpeakerID

O toolkit SpeakerID é uma coleção de scripts (na linguagem Perl) e executáveis (na linguagem C) que permitem a geração e modelagem de características de alto e baixo nível (especialmente, prosódicas). Dentre algumas funcionalidades do toolkit, podemos citar as seguintes:

1. **Detecção de *pitch halving/doubling***: um fenômeno que ocorre durante a produção de fala é a redução (pela metade) da freqüência fundamental no final de frases ou o aumento (dobro) da freqüência fundamental. A detecção de tal fenômeno (o qual também pode ser causado pelo próprio algoritmo de estimação do F0) é de grande importância para a modelagem do locutor como processamento do sinal [Adami, 2004].
2. **Extração de gestos prosódicos**: conjunto de scripts que extraem a seqüência de tokens representando o estado simultâneo dos contornos da freqüência fundamental e energia do sinal de voz (chamado de gestos prosódicos) [Adami, 2004]. O objetivo destas características é modelar as diferenças da produção de fenômenos prosódicos que existem entre os locutores.
3. **Extração de padrões fonéticos**: conjunto de scripts para pré-processamento e geração de seqüência de tokens fonéticos para modelagem de pronúncia ou vocabulário fonético [Andrews et al., 2001]. Duração também é integrada através de quantização da duração do segmento (fonema ou sílaba).
4. **Modelagem de N-gramas**: geração de freqüências de tokens para cálculo de probabilidades condicional ou conjunta dos tokens. O objetivo é localizar padrões específicos para os locutores nos modelos de dependências estatísticas de relevância nas unidades discretas.
5. **Modelagem de misturas gaussianas**: representação do locutor através da combinação linear de várias densidades gaussianas da distribuição das informações. As componentes gaussianas, as quais possibilitam a formação de densidades arbitrárias, representam classes acústicas que caracterizam a voz de um locutor.

Para aproveitar o ambiente multi-processado das redes de computadores, o toolkit oferece scripts para processamento (tanto na extração de características como no processo de reconhecimento) em diversas máquinas. Utilizando o software livre *Sun Grid Engine* – SGE da Sun (http://gridengine.sunsource.net), o toolkit pode realizar processamento paralelo tanto na extração das características como no treinamento e teste dos paradigmas do NIST.

3. Resultados Preliminares

O desempenho do toolkit foi avaliado através do uso do paradigma de avaliação da competição do NIST realizada em 2001 [Martin, 2001]. Os dados para esta competição vêm da base de dados Switchboard I, a qual possui aproximadamente 2400 telefonemas com uma média de 5 minutos de conversação provenientes de todas as regiões dos Estados Unidos. Numa das tarefas da competição, os modelos dos locutores são treinados com aproximadamente 20 minutos de fala. O teste é realizado em um segmento de fala de aproximadamente 2,5 minutos (9990 testes foram realizados).

Um sistema de reconhecimento de locutor baseado no teste da razão de verossimilhança entre o modelo do locutor e um modelo independente do locutor (representando os impostores) [Reynolds et al., 2000] foi desenvolvido utilizando o toolkit SpeakerID. O sistema utiliza o modelo de mistura gaussianas (512 gaussianas para cada modelo) da freqüência fundamental e energia do sinal (e suas respectivas derivadas) para representar cada locutor e os impostores. A decisão de aceitar ou rejeitar a alegada identidade de um locutor é definida através da aplicação de um limiar sobre a razão de verossimilhança entre os modelos do locutor e dos impostores. A Figura 1 mostra as curvas de desempenho dos sistemas (a qual mostra as probabilidades de erro para diferentes limiares) desenvolvido por [Adami, 2004] e desenvolvido utilizando o toolkit SpeakerID. As caixas vermelhas na Figura 1 estão centralizadas no limiar onde a probabilidade de aceitar um impostor (falso alarme) como o locutor legítimo é igual a probabilidade de rejeitar um locutor legítimo (não-detecção). Este limiar produz um erro comumente referenciado como *equal error rate* (EER). Apesar das curvas serem diferentes na Figura 1, tal diferença entre o desempenho do toolkit (EER=14.3%) e do sistema desenvolvido por [Adami, 2004] (EER=15.2%) não é estatisticamente significante [Gillick and Cox, 1989].

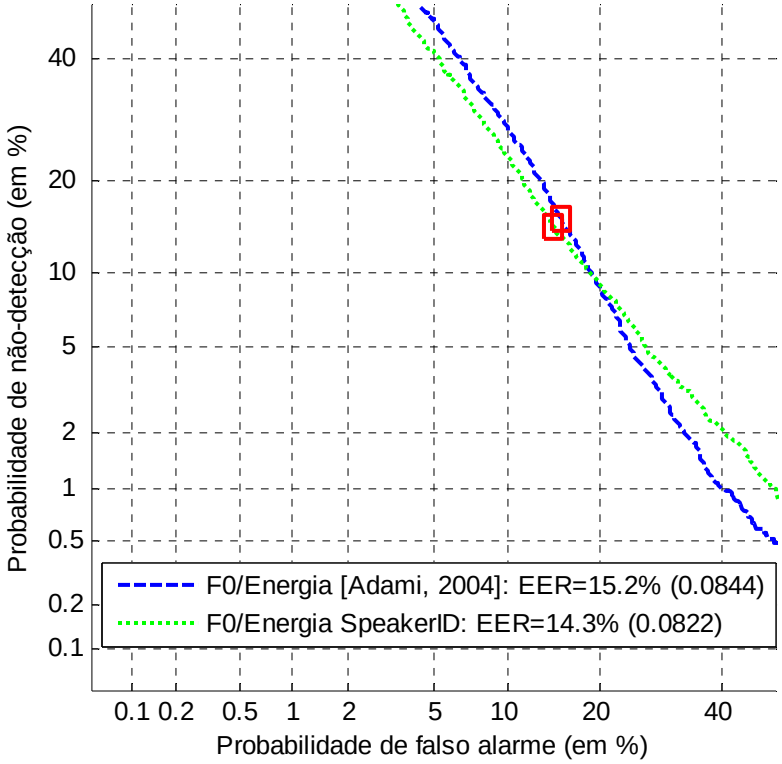

Figura 1. Curvas de desempenho do sistema de reconhecimento proposto em [Adami, 2004] e do toolkit SpeakerID nos dados da competição do NIST 2001.

4. Considerações Finais

Este artigo descreveu um novo toolkit para extração e modelamento de características de alto nível (prosódia e fonética) do sinal da fala. Além disso, tal toolkit permite o desenvolvimento de sistemas de reconhecimento automático do locutor. Atualmente, o toolkit permite a geração de características baseadas em medidas da freqüência fundamental e energia, pré-processamento de seqüências fonéticas, e modelamento de características através das técnicas N-gramas e modelo de misturas gaussianas. O toolkit possui também suporte para a avaliação de sistemas sob o paradigma do NIST. O desempenho de um sistema de reconhecimento de locutor baseado em modelo de misturas gaussianas da distribuição da freqüência fundamental e energia foi apresentado. Foi mostrado que o mesmo obtém desempenho similar ao desempenho publicado em trabalhos anteriores. O toolkit está disponível para download em http://www.ccet.ucs.br/pesquisa/projetos/ralpro/site.

Como trabalho futuro, pretendemos implementar as técnicas de normalização dos escores (do teste da razão de verossimilhança) e de fusão de sistemas. Este trabalho é financiado pelo CNPq, processo número 477845/2004-5.

5. Bibliografia

Adami, A. (2004) Modeling Prosodic Differences for Speaker and Language Recognition. Ph.D. Thesis, OGI School of Science & Engineering at OHSU, Portland, OR, 152 pp.

Andrews, W.D., Kohler, M.A., Campbell, J.P. and Godfrey, J.J. (2001) "Phonetic, Idiolectal and Acoustic Speaker Recognition", 2001: A Speaker Odyssey, Crete, Greece, pp. 55-63.

Bimbot, F. et al. (2004) "A Tutorial on Text-Independent Speaker Verification". EURASIP Journal on Applied Signal Processing, 4: 430-451.

Farahani, F., Georgiou, P.G. and Narayanan, S.S. (2004) "Speaker identification using supra-segmental pitch pattern dynamics", ICASSP, Montreal, Canada, pp. 89-92.

Furui, S. (2005) "50 years of progress in speech and speaker recognition", 10th International Conference on Speech and Computer - SPECOM, Patras, Greece, pp. 1-9.

Gillick, L. and Cox, S.J. (1989) "Some Statistical Issues in the Comparison of Speech Recognition Algorithms", ICASSP. IEEE, Glasgow, Scotland, pp. 532-535.

Hannani, A.E. and Petrovska-Delacrétaz, D. (2005) "Exploiting High-Level Information Provided by ALISP in Speaker Recognition", Non Linear Speech Processing Workshop (NOLISP05), Barcelona, Spain, pp. 19-24.

Lavner, Y., Gath, I. and Rosenhouse, J. (2000) "The Effects of Acoustic Modifications on the Identification of Familiar Voices Speaking Isolated Vowels". Speech Communication, 30: 9-26.

Martin, A. (2001), NIST 2001 Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v05.9.pdf.

Reynolds, D.A., Quatieri, T.F. and Dunn, R.B. (2000) "Speaker Verification Using Adapted Mixture Models". Digital Signal Processing, 10: 19-41.

Schmidt-Nielsen, A. and Crystal, T.H. (1998) "Human vs. Machine Speaker Identification with Telephone Speech", ICSLP, Sydney, Australia, pp. 221-224.