Tecnologia Adaptativa e Síntese de Voz: Primeiros Experimentos

Felipe Augusto Zuffo[1] **, Hemerson Pistori**[1]

[1]CCET - Universidade Católica Dom Bosco
Av. Tamandaré, 6000, Jd. Seminário
CEP: 79117-900 Campo Grande, MS

fazuffo@ec.ucdb.br, pistori@ec.ucdb.br

Abstract. *This paper describes a working in progress that aims at proposing a new methodology for voice-synthesis engineering, based on adaptive technology. The work explores adaptive automata, a Turing-machine powerfull formalism that preserves most of the simplicity of finite state transducers, that has already been used in voice-synthesis for a long time. All the related implementation is being developed with the aid of a free-software, named AdapTools.*

Resumo. *Este artigo descreve um trabalho em fase de desenvolvimento cujo objetivo final é propor uma nova metodologia para geração de sintetizadores de voz, baseando-se em tecnologia adaptativa. Pretende-se explorar, nesse trabalho, os autômatos adaptativos, um formalismo que preserva a simplicidade dos tradicionais transdutores de estados finitos, que já vêm sendo utilizados em síntese de voz, mas que possui o mesmo poder de expressão das máquinas de Turing. A implementação da metodologia será facilitada pela utilização de um software livre para desenvolvimento de autômatos adaptativos, o AdapTools.*

1. Introdução

A síntese de voz consiste na a geração de sinais sonoros que reproduzem as palavras equivalentes em uma determinada linguagem natural. As aplicações que sintetizam a voz são projetadas para imitar a fala, assim como ela é na natureza humana. O crescimento da utilização de sistemas computadorizados pelos mais diversos grupos de pessoas aumenta a necessidade de comunicação ágil entre máquinas e seres humanos. Com o aumento na capacidade de processamento e da disponibilidade de dispositivos para reprodução de som digital, a síntese de voz tem se tornado viável em uma grande variedade de aplicações, como por exemplo, aquelas voltadas para deficientes visuais, em que o monitor é substituído por caixas de som, e aplicações acessadas através de sistemas telefônicos. (e.g. auto-atendimento, telemarketing, etc).

Uma grande quantidade de técnicas vêm sendo exploradas no intuito de otimizar a inteligibilidade da fala sintetizada [Goff and Benoît, 1996, Burrows, 1996, Lemmetty, 1999a]. Para isso, a adequação de fatores como ritmo e intonação são fundamentais para que a saída sonora soe o mais natural possível [Tatham and Lewis, 1996, Lemmetty, 1999a]. Os métodos de síntese de voz são classificados normalmente em três grupos: síntese articulatória, síntese fonética e síntese por concatenação [Lemmetty, 1999b].

Os métodos para síntese articulatória baseiam-se em um modelo do aparelho vocal humano. Mantendo um conjunto de regras que simulam a língua, os lábios e as cordas vocais, o sistema imita a voz humana ao produzir ressonância e articulação. Atualmente, esse tipo de

síntese é pouco utilizado, pois se mostra muito complexa e exige recursos computacionais de alto custo.

As técnicas de síntese fonética são estruturadas sobre um sistema de combinação de freqüências que resulta em uma expressão vocal determinada. A entrada é processada por um conjunto de ressonadores, onde cada um gera a saída equivalente a um som vocal. No final a saída gerada por todos os ressonadores é somada, produzindo um fonema. A disposição dos ressonadores na estrutura de processamento pode variar a cada sintetizador de voz.

Já os métodos de síntese por concatenação, trabalham com uma base de arquivos de som previamente gravados, onde cada fonema é associado a um arquivo. Conforme os fonemas são reconhecidos a partir uma cadeia de entrada, os arquivos são concatenados, de forma a produzir a pronúncia. Com essa abordagem, torna-se mais fácil atingir naturalidade e inteligibilidade nos resultados, porém utiliza-se mais espaço em memória do que nos demais métodos.

Um sistema tradutor texto-voz pode ser definido em dois elementos principais: um mecanismo de reconhecimento de texto e um de geração de som [O'Malley, 1990, Lemmetty, 1999a]. Os dois mecanismos se comunicam através de uma linguagem de representação intermediária, que pode capturar a estrutura das sentenças em níveis morfológicos, léxicos, sintáticos, fonéticos, sub-fonéticos, ou uma combinação de vários níveis. No presente trabalho, em fase de desenvolvimento, estuda-se as possibilidades da utilização de autômatos adaptativos, tanto na fase de reconhecimento do texto, quanto de geração de som. Autômatos adaptativos generalizam o conceito de transdutores de estados finitos (e consequentemente, de autômatos de estados finitos), incluindo características que permitem o tratamento de dependências de contexto, o que os torna um interessante formalismo a ser investigado na solução do problema da síntese de voz. Um software livre, o AdapTools [1], está sendo utilizado nos experimentos com autômatos adaptativos.

Na próxima seção serão apresentados os autômatos adaptativos. Em seguida, descreveremos brevemente o software livre AdapTools. O estado atual do trabalho será relatado na seção 4, seguida das conclusões e metas a serem atingidas posteriormente.

2. Autômatos Adaptativos

Um autômato adaptativo é um dispositivo guiado por regras adaptativo [Neto, 2001] em que a camada subjacente consiste de um autômato de pilha estruturado [Neto, 1987]. Como qualquer dispositivo adaptativo, um autômato adaptativo é capaz de alterar sua estrutura dinamicamente, através de ações elementares que permitem que transições do autômato sejam consultadas, removidas ou inseridas. Essa capacidade de auto-modificação é que torna os autômatos adaptativos mais poderosos que o mecanismo no qual eles se baseiam. Interessantemente, essa capacidade de auto-modificação pode ser facilmente desabilitada, quando o poder extra de expressão não é necessário na solução de um determinado problema.

3. AdapTools

O AdapTools é uma ferramenta livre que oferece um ambiente para experiências, implementação e depuração de autômatos adaptativos [Pistori and Neto, 2003]. O núcleo desse software é uma máquina virtual que executa autômatos adaptativos representados em forma de tabela. Os autômatos e as alterações por ele realizadas são visualizados utilizando-se do pacote livre para tratamento de grafos, OpenJGraph. A figura 1 ilustra a animação gerada pelo AdapTools, de um autômato adaptativo reconhecendo uma cadeia da linguagem não-regular e não-dependente-de-contexto $a^n b^n c^n$.

[1] http://www.ucdbnet.com.br/adaptools/

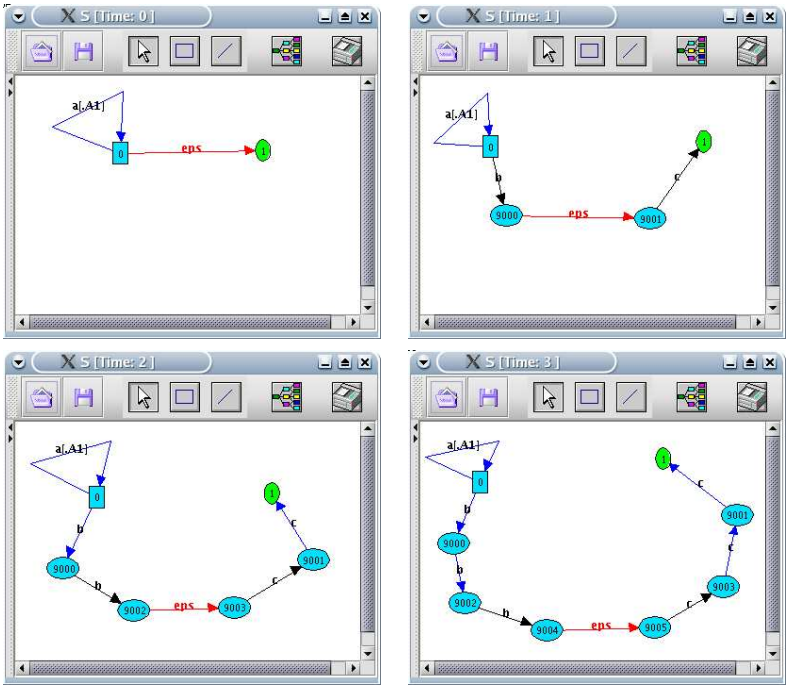

Figura 1: Animação gerada pelo Adaptools para os quatro primeiro estágios de um autômato adaptativo que reconhece $a^n b^n c^n$

4. Síntese de Voz e Autômatos Adaptativos

Utilizando a ferramenta AdapTools, foi desenvolvido um primeiro protótipo de tradutor texto-voz, para a língua portuguesa. A síntese de voz implementada nesse protótipo utiliza um método baseado em concatenação de arquivos armazenados no formato $wave$. A entrada para a tradução texto-voz é uma seqüência de caracteres, que são processados pelo autômato adaptativo, que reconhece sequências fonéticas simples, e mapeia os fonemas reconhecidos em uma rotina semântica capaz de carregar, se necessário, o arquivo $wave$ correspondente ao som, e reproduzi-lo. Para o tratamento de arquivos de som, foram reutilizadas classes disponíveis no pacote livre $SoundPlayer$ [2], criado por Martin Stepp, e que torna a programação utilizando sons extremamente simples. Segue abaixo o trecho de código necessário para reproduzir um arquivo de som utilizando-se o $SoundPlayer$:

```
public SoundPlayer player = new SoundPlayer();

private Token say(String nomeFonema) {
   player.playAndWait(diretorioBase+nomeFonema+".wav");
}
```

A figura 2 mostra algumas das transições do autômato adaptativo que traduz texto em voz. Nas legendas da forma x/y, x indica o símbolo de entrada, e y, o de saída (fonema). As transições $(4, a, 5)$ e $(3, a, 5)$, por exemplo, produzirão ambas o mesmo som (aquele do único fonema da palavra *chá*).

Uma das dificuldades na tradução texto-voz é que uma mesma sílaba, quando encontrada em palavras ou sentenças diferentes, pode ter diferentes pronúncias. Por exemplo, a sílaba xa, nas palavras $fixa$ e $Xuxa$, corresponde a distintos fonemas. Este problema está intimamente

[2] http://www.cs.arizona.edu/~ stepp/java.html

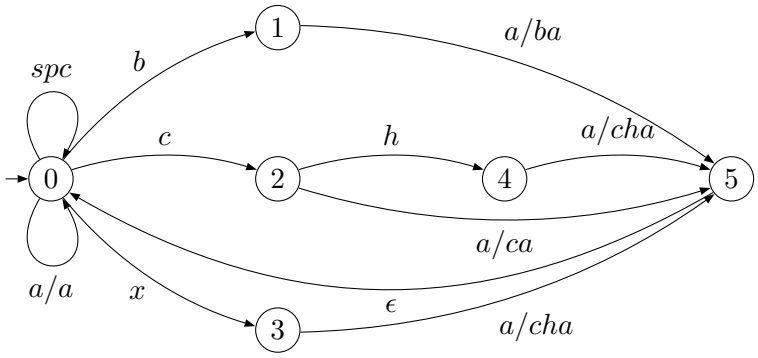

Figura 2: Algumas transições do autômato de tradução texto-voz

ligado à dependência de contexto, o que sugere a utilização de autômatos adaptativos, no lugar de transdutores de estados finitos, na produção de sistemas mais poderosos desta natureza.

5. Conclusão

Apresentamos neste trabalho os primeiros estudos sobre a possibilidade de implementação de tradutores texto-voz utilizando autômatos adaptativos. O fato de a estrutura de um autômato adaptativo poder ser alterada em tempo de execução possibilita a geração de diferentes pronúncias para a mesma sílaba, dependendo do contexto. Essa característica se torna um incentivo para a próxima etapa deste trabalho, o aprimoramento do protótipo apresentado, melhorando a pronúncia de cada palavra. Também pretendemos começar a utilizar um alfabeto fonético internacional, para a representação dos fonemas, e incorporar ao sistema métodos mais eficientes para concatenação de fonemas na formação de palavras.

Referências

Burrows, T. (1996). Speech processing with linear and neural network models.

Goff, B. L. and Benoît, C. (1996). A text-to-audiovisual-speech synthesizer for french. In *Proc. ICSLP '96*, volume 4, pages 2163–2166, Philadelphia, PA.

Lemmetty, S. (1999a). Review of speech synthesis technology. Master's thesis, Helsinki University of Technology.

Lemmetty, S. (1999b). *Review of Speech Synthesis Technology*. PhD thesis, Helsinki University of Technology - Department of Electrical and Communications Engineering.

Neto, J. J. (1987). *Introdução à Compilação*. LTC, Rio de Janeiro.

Neto, J. J. (2001). Adaptative rule-driven devices - general formulation anda case study. In *CIAA'2001 Sixth International Conference on Implementation and Application of Automata*, pages 234–250, Pretoria, South Africa.

O'Malley, M. H. (1990). Text-to-speech conversion technology. *IEEE Computer*, 23(8):17–23.

Pistori, H. and Neto, J. J. (2003). A free software for the development of adaptive automata. *In Proceedings of the IV Workshop on Free Software - WSL (IV International Forum on Free Software)*.

Tatham, M. and Lewis, E. (1996). Improving text-to-speech synthesis. *Proceedings of the Institute of Acoustics*, 18(9):35–42.