RECONHECIMENTO DE VOZ PARA O PORTUGUÊS BRASILEIRO

Aluno: Jan Krueger Siqueira
Orientador: Abraham Alcaim

Introdução

Uma das questões atuais de acessibilidade é a criação de sistemas capazes de reconhecer e eventualmente executar alguma ação a partir da voz humana. Entre suas inúmeras aplicações, podemos citar a interação de deficientes visuais com o computador e a operação de equipamentos quando um trabalhador não se encontra com as mãos disponíveis.

Embora já existam inúmeras pesquisas e produtos nessa área, poucos deles são compatíveis com o idioma português do Brasil.

Objetivos

Estudar os princípios e recursos do reconhecimento de voz. Preparar um banco de vozes para treinar o sistema de reconhecimento. Pesquisar em artigos a utilização e os resultados de diversos sistemas de reconhecimento. Configurar os parâmetros do reconhecedor. Testar o sistema.

Metodologia

Existem inúmeros parâmetros que caracterizam a voz humana: espectro de freqüência, número de cruzamentos com zero, análise do sinal e suas derivadas no domínio do tempo, coeficientes cepestrais, etc. Tais parâmetros podem ser usados, por exemplo, na compressão de voz para transmissões digitais ou para arquivos de áudio. Outra de suas aplicações ocorre em diversos sistemas de reconhecimento, entre eles o muito bem sucedido Modelo de Markov Escondido (Hidden Markov Model – HMM). Dessa forma, pensou-se na idéia do reconhecimento remoto a partir de voz obtida de celulares ou de voz sobre IP.

Existem diversos tipos de codificadores utilizados na telefonia móvel e IP atualmente. Para serem usados no sistema de reconhecimento, serão inicialmente testados os seguintes: AMR-NB, AMR-WB e G723.1. Todos eles possuem código fonte aberto, podendo ser compilados em diversos sistemas operacionais.

Após a leitura das especificações dos três codificadores, foi possível utilizá-los na preparação e conversão do banco de vozes. Já estava disponível nesse banco um total de 1000 frases, cada uma pronunciada por 100 locutores (50 homens e 50 mulheres), num total de 100.000 (cem mil) arquivos em formato ".wav" (mono, 16 bits, 16 kHz).

Adequar tais arquivos à entrada dos codificadores exige conhecimento da estrutura do formato "wav" e manipulação em massa de bytes. Já a conversão automática de todo o banco de vozes demanda acesso ao sistema de pastas e arquivos do sistema operacional, bem como a invocação em loop da execução dos codificadores (via prompt). Considerando tudo isso, a ferramenta de apoio que se mostrou mais apta e prática foi o MatLab.

De um modo geral, as modificações nos arquivos "wav" não são muito complexas: os codificadores AMR-NB e G723.1 exigem um arquivo amostrado a 8 kHz, fazendo-se necessária uma filtragem e um downsampling; o AMR-WB processa amostras de 14 bits e o AMR-NB processa com 13 bits, o que requer a anulação e aproximação de bits mais significativos; e o cabeçalho com os metadados é sempre retirado. Esses processamentos geram um arquivo ".inp".A compilação dos codificadores também não toma muito tempo:

escritos na linguagem C, podem ser transformados em executáveis para Windows ou Unix via GCC. Utilizam-se então, como teste, alguns arquivos com os quais se converte do formato ".wav" para ".inp", e os codificadores e decodificadores para processar os arquivos resultantes. Finalmente, adapta-se o resultado de volta para ".wav". Comparando-se o som original com o final, pode-se verificar a funcionalidade do sistema.

Em seguida, a codificação em massa foi testada. Para organizar os dados, o banco de vozes foi catalogado em pastas separadas de acordo com o locutor. A programação garantiu que os arquivos convertidos fossem organizados do mesmo modo, porém num diretório diferente dos originais.

A primeira etapa do projeto foi então concluída. No momento, pesquisa-se em artigos os resultados e conclusões sobre esta linha de reconhecimento de voz envolvendo codificadores. Outro ponto a ser investigado é o funcionamento mais detalhado dos codificadores/decodificadores, já que a intenção é alterá-los para que forneçam os parâmetros mais adequados ao reconhecedor.

Próximos Passos

Utilizando o HTK (HMM Tool Kit), o reconhecedor de voz será ajustado para os fones e a gramática brasileira. O treinamento será realizado com os parâmetros de voz obtidos por cada um dos codificadores, na busca de saber qual fornece a melhor taxa de reconhecimento.

A expectativa é que o sistema seja capaz de identificar a fala contínua, independentemente da escolha do locutor.

Referências

1 – RABINER, L.; JUANG B. **Fundamentals of Speech Recognition**. New Jersey, USA: Prentice-Hall, c1993. 507p.

2 – RABINER, L. A tutorial on hidden Markov models and selected applications in speech recognition. **Proceedings of the IEEE**, v. 77, n. 2, pp. 257-286, fev. 1989.

3 – http://wiki.multimedia.cx/index.php?title=AMR-NB

4 – http://en.wikipedia.org/wiki/G.722.2

5 – http://www.vocal.com/data_sheets/g723.html

6 – http://htk.eng.cam.ac.uk/