1
Introdução

A área de reconhecimento de voz [1] busca desenvolver sistemas que compreendam a fala humana, ou seja, que reconheçam aquilo que foi dito por uma pessoa. Com isso, ações podem ser efetuadas de modo verbal, facilitando o trabalho do usuário (por exemplo, nos momentos em que suas mãos não podem ser usadas). Ao longo dos anos, esses sistemas se desenvolveram bastante e agora estão presentes em diversas aplicações: ditado de textos, atendimento automático por telefone, auxílio a deficientes físicos, etc.

O grau de complexidade de um reconhecedor de voz varia de acordo com a quantidade de palavras a serem usadas. Por exemplo, a identificação de um comando do tipo sim/não é mais simples do que a de uma sequência de dígitos numéricos, que por sua vez é mais simples do que a de frases variadas. Da mesma forma, a dificuldade aumenta se o sistema tiver que funcionar com qualquer locutor, ao invés de ser direcionado a apenas um. Logo, o caso de frases pronunciadas por qualquer pessoa – chamado de reconhecimento de voz contínua (RVC) independente de locutor – é o cenário mais crítico e também o escolhido para este trabalho.

Um dos maiores desafios na área de RVC é a presença de ruído. No momento da gravação da voz, perturbações sonoras como burburinho ou ronco de motores podem ser capturadas também. Assim, a informação das palavras fica corrompida, o que prejudica bastante a identificação do que foi dito.

Portanto, o objetivo deste trabalho é analisar e testar técnicas que diminuam os problemas causados pelo ruído. Muitos artigos sobre esse tema foram publicados, mas os autores quase sempre focaram em apenas uma abordagem. Aqui, três técnicas serão combinadas entre si para gerar um sistema mais robusto.

A primeira técnica é a extração de atributos (informações) da fala. Na literatura, existem vários métodos propostos. Um deles, chamado de *Mel-Frequency Cepstral Coefficients* (MFCC) [2], ainda é muito utilizado, porém seu desempenho cai rapidamente com a presença do ruído. Em [3], o método *Subband*

Spectral Centroid Histograms (SSCH) foi introduzido e gerou resultados melhores que o MFCC. Mais recentemente, [4] apresentou os atributos *Power-Normalized Cepstral Coefficients* (PNCC), também mais robustos que o MFCC. Esses três métodos serão avaliados e comparados entre si.

A segunda técnica é o *wavelet denoising* [5]. Através da transformada *wavelet*, a informação do ruído é filtrada, resultando num sinal de voz menos corrompido. O procedimento usa uma função de limiar, que é simples e eficaz.

A terceira e última técnica é uma proposta original deste trabalho, chamada de *feature denoising*. Após a extração de atributos, os valores produzidos são inseridos em redes neurais [6], gerando novos valores menos corrompidos. Trata-se de um novo tipo de remoção de ruído, agora realizada entre a extração e o reconhecimento.

As três técnicas citadas são combinadas em etapas separadas, como mostra a Figura 1. Cabe ressaltar que o sistema proposto traz uma série de resultados inéditos. Por exemplo, não foram encontradas referências de testes dos métodos SSCH e PNCC com o *wavelet denoising*, nem mesmo uma comparação entre ambos.

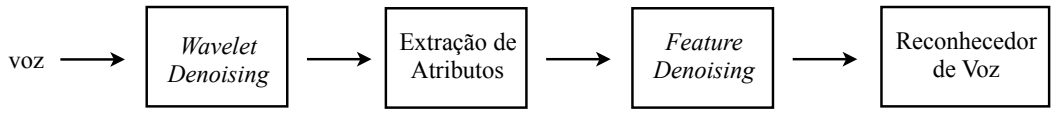

Figura 1: Etapas do sistema proposto para reconhecimento de voz.

Conceitos básicos sobre reconhecimento de voz são explicados no capítulo 2. Métodos de extração de atributos são apresentados no capítulo 3. A técnica de *wavelet denoising* é descrita no capítulo 4. O *feature denoising* é proposto no capítulo 5. O detalhamento do experimento prático e seus resultados são vistos no capítulo 6. Conclusões e sugestões para trabalhos futuros são levantadas no capítulo 7. E, finalmente, os apêndices de A até F contêm demonstrações dos algoritmos utilizados.