7
Conclusões e Sugestões para Trabalhos Futuros

A grande motivação para o presente trabalho foi a de propor técnicas que fossem capazes de superar alguns dos obstáculos encontrados para que os reconhecedores de voz contínua distribuída com amplo vocabulário tenham um bom desempenho.

7.1.
Conclusões

A construção da base de vozes (base Alcaim – Alencar) foi a primeira contribuição relevante para o desenvolvimento de sistemas de reconhecimento de voz contínua para o Português Brasileiro obtido nesta tese. A Base de voz construída visa o treinamento e teste de sistemas de reconhecimento de voz contínua para o Português Brasileiro com amplo vocabulário e independentes do locutor (100 locutores – 50 locutores do sexo masculino e 50 locutores do sexo feminino, cada um falando todas as 1000 frases foneticamente balanceadas escolhidas para compor a base [2]). A gravação só foi possível devido ao apoio do CNPq (através de projeto aprovado em edital Universal) que permitiu a contratação da Audioteca Sal e Luz (ONG que visa a inclusão de deficientes visuais através da gravação de livros falados) que gravou a base em seus estúdios, obteve os locutores e forneceu os arquivos para avaliação e inserção na base. Os locutores desta base têm idades variando de 17 à 65 anos, sendo alguns deles profissionais na realização de locuções, outros treinados para a gravação de livros falados e outros sem nenhum conhecimento teórico ou prático sendo orientados durante a gravação. A verificação de todos os arquivos de áudio para garantir que o padrão em qualidade, nomenclatura, frase lida, taxa de bits, etc, havia sido respeitado, demandou um grande esforço de paciência e tempo, devido ao tamanho da base e para garantir a qualidade dos resultados a serem obtidos.

A gravação foi realizada em estúdio, ambiente sem ruído, com uma especificação de gravação que pudesse abranger a entrada dos diversos

codificadores de voz utilizados em Telefonia Móvel Celular e IP (taxa de amostragem 16 kHz e 16 bits por amostra com banda de sinal de 50 – 7000 Hz.).

Para um bom funcionamento dos Sistemas Automáticos de Reconhecimento de voz é necessário que os atributos de reconhecimento sejam obtidos a uma taxa elevada, porém os codificadores de Voz para Telefonia IP e Móvel Celular, usados em cenários distribuídos, normalmente geram seus parâmetros a taxas mais baixas, o que degrada o desempenho do reconhecedor. Usualmente é utilizada a interpolação linear no domínio das LSFs para resolver este problema. Nesta Tese foi proposta a realização da interpolação com a utilização de um Filtro Digital Interpolador que demonstrou ter um desempenho de reconhecimento muito superior ao da interpolação linear.

No primeiro conjunto de testes para o ITU-T G.723.1, foram realizados alguns experimentos importantes para o cenário de reconhecimento de voz contínua distribuído em amplo vocabulário para o Português Brasileiro. Foi mostrado que apenas a independência do locutor deteriora em torno de 4% a taxa de reconhecimento. Adicionalmente, uma redução de 6% no desempenho foi observada pelo uso de diferentes frases no treinamento e testes do sistema. Mostrou-se também que a MFCC, obtida de voz reconstruída, é bastante sensível ao ruído de codificação, reduzindo o desempenho de 14% aproximadamente. Foi possível observar também que a MFCC de voz reconstruída tem uma sensibilidade em torno de 2% pela forma que a excitação é codificada e decodificada (comparando a linha dois – operação em 5,3 kbits/s – e a linha três – operação em 6,3 kbits/s – de cada tabela do capítulo 5). Os atributos obtidos dos parâmetros LSF ou LPC podem prover um melhor desempenho do que os obtidos da voz reconstruída (MFCC). O MPCEP é o melhor atributo para ser utilizado em um sistema LVDCSR (*Large Vocabulary Distributed Continuous Speech Recognition*) empregando o codificador ITU-T G.723.1. O mesmo oferece uma melhor $\overline{WRR}$ (*Average Word Recognition Rate*) com menor complexidade.

Para o segundo conjunto de testes mostrou-se que a interpolação usando filtros digitais das LSFs decodificadas melhora significativamente o desempenho de todos os atributos de reconhecimento obtidos do codificador ITU-T G.723.1 quando comparados com a interpolação linear. O $\overline{WRR}$ melhorou de

aproximadamente 4% em todas as situações onde os atributos são obtidos de parâmetros LSF e LPC.

Comparando o ITU-T G.723.1 nas taxas 6,3 e 5,3 o kbits/s com o codec AMR-NB nas taxas similares (6,7 e 5,9 kbits/s) a última fornece um ganho de 1.55% na $\overline{WRR}$. É importante notar que os intervalo de confiança de 95% nestes dois casos não se sobrepõem. É também importante observar que o AMR-NB em suas taxas mais baixas (5,15 e 4,75 kbits/s) supera o codec de ITU-T G.723.1, que opera em 6,3 e 5,3 kbits/s. Além disso, seus intervalos de confiança de 95% outra vez não se sobrepõem. Deve ser anotado que as LSFs no AMR-NB são codificadas em uma taxa de bits mais elevada do que a usada pelo ITU-T G.723.1.

Para o terceiro conjunto de testes evidenciou-se que os ISFs do codificador AMR-WB são inadequados para uso como atributo de reconhecimento de voz, tendo seu desempenho sido superado inclusive pela MFCC de voz reconstruída.

O problema de perda de pacotes em rajadas nas redes IP e redes móveis é um dos fatores mais importantes a serem considerados na análise de sistemas de reconhecimento de voz distribuídos. Perdas de pacotes em rajadas causam uma redução drástica do desempenho do reconhecimento de voz.

Existem algumas aproximações para melhorar o desempenho do sistema de reconhecimento de voz na presença de imperfeições do canal, tais como apagamentos dos quadros. Uma solução simples é inserção dos zeros na posição dos pacotes perdidos. Uma outra aproximação é interpolação linear, entre pacotes recebidos com sucesso (em nosso caso, quadros). O destino recebe, por exemplo, o primeiro conjunto de LSFs quantizadas. Entretanto, devido às imperfeições do canal, não é recebido o segundo conjunto. Na chegada do terceiro conjunto, o receptor pode aproximar o segundo pela interpolação linear do primeiro conjunto com o terceiro. Certamente, a interpolação de mais de um conjunto é praticável em troca de um incremento indesejável de atraso [80]. Para aplicações de redes IP, se n quadros consecutivos de duração t cada um, é perdido, o atraso devido à interpolação é $D_i = nt + RTT/2$, onde RTT (*Round-Trip Time*) é o tempo para um pacote ir da fonte ao destino e então de volta à fonte. Anote que valores típicos para RTT variam de 10 a 700 ms e de acordo com [80], atrasos aceitáveis para aplicações de VoIP não devem exceder 800 ms.

É importante notar que a primeira técnica (Inserção Zero) ignora as características do sinal. Consequentemente, não explora o conhecimento do sinal para melhorar o desempenho do reconhecimento. Por outro lado, o uso da segunda técnica (Interpolação Linear) implica geralmente em longo atraso nos pacotes reconstruídos.

Foi proposta uma técnica nova para a reconstrução dos pacotes perdidos baseada em Redes Neurais e comparou-se seu desempenho de reconhecimento com os aqueles obtidos com as técnicas de inserção de zeros e interpolação linear.

Dos resultados da simulação fica claro que a Inserção de Zeros é definitivamente a pior aproximação para a solução da perda de pacotes. Comparando a Inserção de Zeros, a Interpolação Linear e a Redes Neurais para a reconstrução de pacotes perdidos de LSFs, pode-se ver que a técnica proposta que usa Redes Neurais supera as duas outras técnicas em todos os casos. Entretanto, as melhorias são somente significativas nas perdas de pacotes - *TPP* - de 30% e 40%, respectivamente, onde as condições das redes IP e das redes móveis são mais severas. No caso onde $TPP = 40\%$, o esquema novo fornece ganhos de reconhecimento de aproximadamente 3% quando comparado com a técnica da Interpolação Linear.

Agora comparando o ITU-T G.723.1 nas taxas 6,3 e 5,3 kbit/s com o codec de AMR-NB nas taxas similares (6,7 e 5,9 kbit/s), em toda as condições de rede, o AMR-NB fornece um ganho em torno de 1,50%. É também muito significativo notar que os intervalos de confiança de 95% não se sobrepõem. Além disso, é importante observar que o AMR-NB em suas taxas mais baixas (5,15 e 4,75 kbit/s) supera o codec de ITU-T G.723.1 que opera em 6,3 e 5,3 kbit/s para todos os valores de *TPP*. Outra vez, seus intervalos de confiança de 95% não se sobrepõem. Note que as LSFs de onde os atributos de reconhecimento são extraídos, são codificados em uma taxa de bits mais elevada pelo AMR-NB em comparação ao ITU-T G.723.1. Finalmente, ficou claro que as Redes Neurais representam uma técnica atrativa para a reconstrução de pacotes perdidos para ambos os codificadores de voz.

7.2.
Sugestões para Trabalhos Futuros

Uma primeira sugestão interessante seria obter matematicamente e testar atributos de reconhecimento para o codificador AMR-WB diretamente dos parâmetros ISFs, que eventualmente apresentassem um desempenho semelhante ou superior ao obtido com os atributos obtidos de LSFs.

Fica também como sugestão, para trabalhos futuros, a busca por técnicas que possam melhorar o desempenho de sistemas de reconhecimento de voz distribuído na presença de ruído ambiente.

Uma experiência interessante a ser realizada consiste em verificar o comportamento de atributos robustos ao ruído, como por exemplo o ZCPA [10], quando obtidos de voz reconstruída por um decodificador padrão e compará-los com os melhores atributos aqui obtidos.

Um outro caminho a ser validado para a melhoria do desempenho do decodificador é a fusão de diferentes atributos de reconhecimento que possam de alguma forma agregar informações diferentes sobre a voz a ser reconhecida.