Reconhecimento de palavras faladas isoladas em dispositivos móveis

Raphael F. Ramos, Luis A. Rivera, Fermín A. Tang, Ausberto S. Castro

raphaelframos@gmail.com, {rivera, ascv, tang}@uenf.br

Laboratório de Ciências Matemáticas – LCMAT
Universidade Estadual do Norte Fluminense – UENF
Av. Alberto Lamego, 2000, CEP 28015-620 Campos dos Goytacazes, Rio de Janeiro – Brasil

Resumo: *A fala é a principal forma de comunicação dos seres humanos. Ao falarmos permitimos ao outro que conheça nossos pensamentos, sentimentos, necessidades e passamos a conhecer os sentimentos, pensamentos e necessidades do outro. Essa forma de comunicação humana ainda é um sonho se tratando de comunicação homem-máquina. O fator motivador para pesquisa nesta área foi a possibilidade de desenvolvimento de um sistema de reconhecimento de fala que não necessitasse de algum tipo de aprendizado para sua utilização, possibilitando que qualquer pessoa através da fala pudesse manipular o sistema, já que a fala é o meio de comunicação mais natural para o ser humano. Esse trabalho visa construir um jogo para descoberta de cores utilizando a voz para o seu reconhecimento, este sistema servirá como base para trabalhos futuros sobre reconhecimento de fala já que as técnicas mais tradicionais nesses sistemas foram aplicadas. O reconhecedor desenvolvido é aplicado ao reconhecimento de palavras isoladas com dependência de locutor que roda em dispositivos móveis com sistema operacional Android.*

Palavras chaves: Reconhecimento de voz, Fala isolada, Interação, Comando voz, Móvel.

Abstract: *Speech is the primary form of communication between human beings. By talking to others we allow them to know our thoughts, feelings, needs and also come to know their feelings, thoughts and needs. This form of human communication is still a dream if we compare it to man-machine communication. The motivating factor for research in this area was the possibility of developing a speech recognition system that did not require any type of learning for their use, allowing that anyone could manipulate the system through speech, since the speech is the more natural medium for human communication. This work aims to build a game for discovery of colors using speech for recognition, this system will serve as a basis for future work on speech recognition as the more traditional techniques were applied in these systems. The recognition system developed was applied to the recognition of isolated words with speaker dependency and runs on mobile devices with Android operating system.*

Keywords: Voice recognition, Speech isolated, Interaction, Voice command, Mobile.

1 Introdução

Desde os inicios dos computadores, a fala foi considerada como desafio de comunicação humano-computador. Alain Turing, na época, propôs o "Test de Turing" como o desafio na evolução da computação que a comunicação entre humano e computador seja na língua falada. Mas isso, possivelmente, nunca seja superado devido aos muitos fatores humanos envolvidos no processo da comunicação humana. Um segmento da comunicação por fala é *reconhecimento automático de fala*– RAF (do inglês *automatic speech recognition*, ASR), que consiste em que um sinal de som capturado pelo microfone seja analisado e conferido se ele representa um elemento de comunicação oral, tal como uma palavra específica de um domínio.

Os sistemas RAFs são atualmente alvo de pesquisas com diversos enfoques: identificação do falante, controle de iteração com dispositivos móbiles, jogos, aplicativos em tem real, no comercio, na telefonia, etc. Por exemplo, as demandas de crianças em jogos que permitam uma interação por voz é exigência do mercado; por outro lado, estão os adultos cada vez exigentes com os móveis demandando uma interação natural por voz, por exemplo, quando estão realizando alguma atividade manual, queiram fazer uma consulta por nomes da agenda, ou por número de telefones parciais, por endereços, marcar números, etc.

Nos aspecto técnico, se busca técnicas de forma que possam atender essas necessidades humanas, considerando as limitações do sistema hardware para armazenamento e velocidade de processamento dos móveis, porque as técnicas de reconhecimento da fala, em geral, demandam processamentos de grandes quantidades de informações de sinais de som, e em várias etapas, que comprometem a capacidade de resposta eficiente e em tempo real.

Neste trabalho formula-se um modelo de reconhecimento de fala isolada com pouca demanda de espaço de armazenamento e sem comprometer a velocidade de resposta no processamento de reconhecimento da fala.

O trabalho organiza-se da seguinte forma: na Seção 2 abordam-se reconhecimento automático de falas. Na Seção 3 se formula o modelo de reconhecimento de fala isolada para móbiles; na Seção 4 aborda-se a representação de conhecimento e reconhecimento; na Seção 5 detalha-se a implementação e se discutem os resultados. Finalmente, na Seção 5 conclui-se coma indicação de trabalhos futuros.

2 Sistemas de Reconhecimento Automático de Fala

Os sistemas de reconhecimento automático de fala têm como objetivo transformar um sinal analógico (som emitido pelo falante) obtido através de um transdutor em

uma sequência de fonemas que compõem uma fala. Normalmente, um sistema RAF é dividido em quatro fases: aquisição, pré-processamento, extração de informações, e classificação.

Na fase de *aquisição*, as ondas sonoras são convertidas em sinais elétricos que vão ser representados por números. No processo de captura, geralmente, ocorrem perturbações de sinais devido às características do ambiente, como presencia de ruído, variações de sinais, etc. Na segunda fase, *pré-processamento*, os sinais elétricos são purificados do ruído a fim de tornar o sinal o mais próximo possível da fala original, removendo os períodos de silencio, normalizando o volume da elocução e estruturando em subsequências de sinais de fala, conhecido como janelas, para uma melhor caracterização. Na terceira fase, que é a fase de *extração de informações* (características), é obtido um menor número de parâmetros que representa a informação toda, que neste trabalho será chamado de vetor de referência, de forma que na seguinte fase sejam realizadas as *operações de classificação*. A Figura 1 ilustra o esquema de um sistema de RAF com as fases mencionadas.

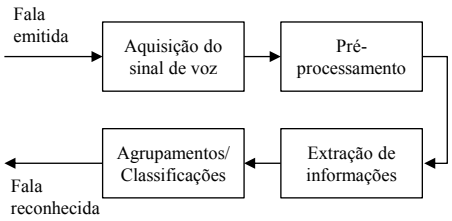


Figura 1: Esquema geral de um sistema de RAF.

Segundo esse esquema, utilizando diferentes abordagens de implementação, são processadas as diferentes categorias de falas.

2.1 Categorias de RAFs

Existem vários enfoques para categorizar um RAF, e as mais importantes, segundo Lee (1989), são: dependência do locutor, modo de pronúncia, tamanho do vocabulário, perplexidade, e relação sinal-ruído.

Desde o ponto de vista de *dependência do locutor*, os RAFs são classificados como dependentes e independentes do locutor. No dependente do locutor o sistema só reconhece a fala das pessoas cujas vozes foram utilizadas no treino. Enquanto no tipo independente do locutor o RAF reconhece a palavra emitida por qualquer pessoa. Nesta forma de classificação, a implementação dos sistemas independentes do locutor é mais simples porque o treino é realizado com falas de diferentes pessoas, de diferentes idades, sexo, sotaques, etc.

No enfoque *modos de pronúncia,* os sistemas RAF podem ser classificados de duas formas: sistemas de palavras isoladas e sistemas de falas conectadas. Os *Reconhecedores de palavras isoladas* são sistemas que reconhecem palavras faladas isoladamente; isto é, entre cada palavra deve existir uma pausa mínima, para que seja detectado o início e o fim da mesma. Isso proporciona um resultado muito superior aos da fala conectada, esses sistemas são os mais simples de serem implementados. Os *reconhecedores de palavras conectadas* são sistemas mais complexos que utilizam palavras como unidade fonética padrão. São capazes de reconhecer sentenças completas, pronunciadas sem pausa entre as palavras, e por não ter informação de onde começa e terminam determinadas palavras, muitas delas são mascaradas, encurtadas e às vezes não pronunciadas (Zue, 1985). Esses sistemas precisam lidar com todas as características e vícios da linguagem natural, como o sotaque, a duração das palavras, a pronúncia descuidada, etc.

No enfoque *tamanho do vocabulário*, quanto maior é seu tamanho, maior será a quantidade de palavras ambíguas, com realizações sonoras semelhantes, ocasionando maior chance de erros por parte do decodificador responsável pelo reconhecimento. Em (Silva, 2009) os vocabulários podem ser definidos como pequenos (até 20 palavras), médios (entre 20 a 100 palavras), grandes (entre 100 a 1000 palavras) e muito grandes (mais de 1000 palavras). Sistemas que reconhecem grandes quantidades de vocabulários são chamados de *large vocabulary continuos speech recognition (LVCSR)*.

No enfoque *Perplexidade* prima o modelo de domínio de discurso em uma linguagem. Então, perplexidade classifica os sistemas RAFs em relação à quantificação da dificuldade de comparações que as linguagens impõem. Uma métrica que mede a dificuldade da tarefa, combinando o tamanho do vocabulário e o modelo de linguagem, é definida como a média do número de palavras que pode conseguir uma palavra depois que o modelo de linguagem for aplicado.

Finalmente, no enfoque *relação sinal-ruído*, também conhecido como *Signal Noise Ratio* (SNR), um sistema RAF é classificado considerando os problemas que podem prejudicar o seu desempenho, tais como ruídos, ambiente, distorção acústica, diferentes microfones e outros.

2.2 Trabalhos relacionados

Como os sistemas RAFs são importantes para os seres humanos, muitos estudos foram feitos no sentido de utilizar a fala para realizações de ações nos sistemas computacionais. Essas aplicações são dadas em várias áreas de necessidade humana; por exemplo, um sistema controle de cadeira de rodas através de comandos de fala, desenvolvido por Barcelos (2007), útil para cadeirantes com problema de mãos e braços. Nesse trabalho foi utilizado o software IBM *Via Voice*, que segundo Damasceno (2005) obteve um melhor desempenho e aplicabilidade quando comparado aos outros softwares, considerando a língua falada, a robustez do reconhecimento e a usabilidade da interface. Outro trabalho apresentado por Bresolin (2003) consiste em um sistema de RAF para comandar um equipamento elétrico qualquer, e foi implementado utilizando as bibliotecas de MATLAB.

Rodrigues (2009) utiliza um sistema de redes neurais artificiais (RNAs) que identifica comandos de voz para acionar um robô móvel. Uma RNA verifica as características vocais de um locutor e reconhece a fala desse locutor. O sistema foi desenvolvido como dependente do locutor. Nesse modo, para cada novo

locutor é necessário um novo treinamento da rede com as características vocais desse locutor. Em outro trabalho, Paula (2000) utiliza também as RNAs para a criação de um sistema de RAF de palavras faladas na língua portuguesa como: “um”, “dois”, “três”, etc.

Ruaro (2010) apresenta um RAF para dispositivos móveis, implementado em *java*, utilizando a técnica de deformação dinâmica no tempo (DTW: *dynamic time warping*) para reconhecimento de pronuncias em português dos dígitos de 0 a 10. Ramiro (2010) desenvolve um sistema RAF, utilizando HMM (*Hidden Markov Model*), para palavras isoladas como “ligar” e “desligar” uma “televisão” ou uma “lâmpada”. Louzada (2010) apresenta um sistema de RAF que só realiza os comandos ditos com mais de 70% de acerto no reconhecimento utilizando HMM para um sistema independente do locutor. Outro sistema RAF, independente de locutor, é proposto por Alvarenga (2012) para identificar comandos de voz para controlar movimentos de robôs, com aplicações na indústria e no auxílio de deficientes físicos.

3 Modelo de RAF para móveis

O modelo do sistema de RAF proposto é um sistema independente de locutor, que consta de um conhecimento envolvendo um grupo de padrões e processamento de reconhecimento implementado em *smartphones* com sistema operacional *Android*. Cada usuário precisa treinar o sistema para conseguir uma melhor eficiência no reconhecimento. Utiliza-se o modo de pronúncia de palavras isoladas, onde cada amostra a ser reconhecida é salva no dicionário do sistema. Para verificar o modelo utiliza-se um dicionário pré-definido com as palavras: “Amarelo”, “Azul”, “Branco”, “Preto”, “Verde” e “Vermelho”. O reconhecimento é realizado comparando a elocução teste com esses padrões.

O sinal sonoro de cada uma dessas palavras isoladas é capturado pelo microfone do celular e convertido em sinal digital. Seguidamente, aplica-se ao sinal digital uma serie de filtros, tais como: detecção de extremos, normalização, retirada do nível DC, pré-ênfase e janelamento. Na extração de informação, são removidas as informações redundantes através do método *Mel-frequency cepstrum* (MFC), com esse resultado obtêm-se as características para formar padrões de classificação. Os padrões estão definidos em torno das falas presentes no dicionário. Os diferentes sinais digitais associados às falas, em sua forma de representação de características, são comparados em função das distâncias, utilizando a técnica conhecida como DTW (*Dynamic Time Warping*). Assim, finalmente, uma fala reconhecida, e com ajuda de um dicionário de cores permitidas, é mostrada a cor correspondente na tela. Com essas operações, para propósito deste trabalho, a arquitetura proposta é ilustrada na Figura 2.

3.1 Aquisição da fala

A *aquisição do sinal de fala* é responsável pela captura e conversão do sinal analógico de uma voz, emitido por um falante, em um sinal elétrico. As ondas sonoras produzidas pelo falante são capturadas como sinal analógico por um aparelho transdutor, e se converte em sinais elétricos. A distância entre o falante e o transdutor, como ilustra a Figura 3, é um dos fatores que as ondas sonoras primárias sejam contaminadas por ruídos do ambiente.

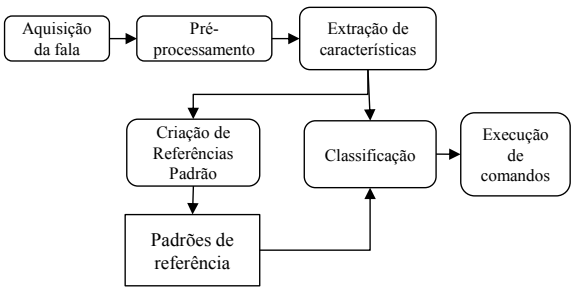


Figura 2: Arquitetura do modelo RAF abordado.

Na aquisição em dispositivos móveis a captura é feita pelo microfone do celular ou *tablet*. Nessa etapa realiza-se a filtragem do sinal capturado por um filtro de passa-baixa, chamado *anti-aliasing*. Esse filtro tem o intuito de suprimir componentes de frequências superiores à metade da frequência de amostragem, como exigido pelo teorema de Nyquist (Farias, 2011).

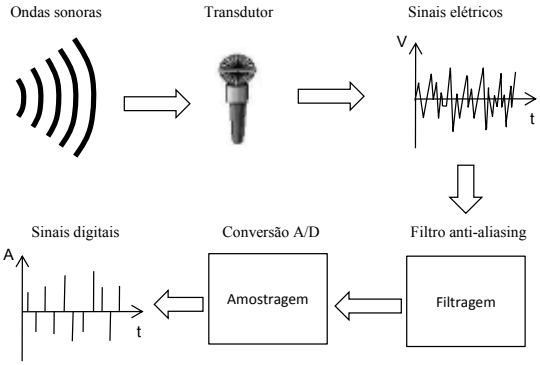


Figura 3: Processo de aquisição do sinal da fala.

3.2 Pré-processamento

O sinal digitalizado, apesar da filtragem de *anti-aliasing* realizada, arrasta ruídos de alta frequência, períodos de silencio, etc. O objetivo desta etapa é deixar a informação consistente e simples de operar na seguinte etapa. Para isso, o sinal digital é submetido às operações, tal como ilustra a Figura 4, de detecção de extremos, normalização de amplitude, retirada de nível DC e janelamento.

A informação digitalizada, no formato PCM (*Pulse-Code Modulation*), contem um cabeçalho que deve ser removido para melhorar o desempenho das operações.

Na detecção de extremos, o sinal digital é analisado para a identificação do inicio e fim da locução, para remover os períodos de silêncio que podem conter ruídos, sinais indesejados, e, também, encontrar a duração do sinal falado. Esse processo, também, permite diminuir a carga computacional e economizar tempo no processo apenas de trechos realmente da fala (Chu, 2003). O *extremo inicial* é determinado pelo primeiro quadro onde realmente se inicia a fala, e o *extremo final* é determinado pelo último quadro que ainda há fala. O algoritmo utilizado neste caso, apresentado em (Saha et al, 2005), consiste em: primeiro, calcular a média M e o desvio padrão das primeiras 1600 amostras do vetor

normalizado; segundo, em cada amostra verificar se a função distância de Mahalanobis (Sena et al, 2010) é maior do que 3, então marcar a amostra com 1 porque é uma amostra sonora, caso contrário marcar com 0 porque é silêncio. Assim se tem um vetor de 0s e 1s. Estabelecer segmentos de 10ms como janelas, na janela que tiver maior número de 0s que os 1s converter em 0s, caso contrario converter em 1s. Em um processo seguinte coletar a parte sonora, aqueles rotulados por 1s, em um novo vetor que vem a ser o vetor do sinal digital como silêncio removido.

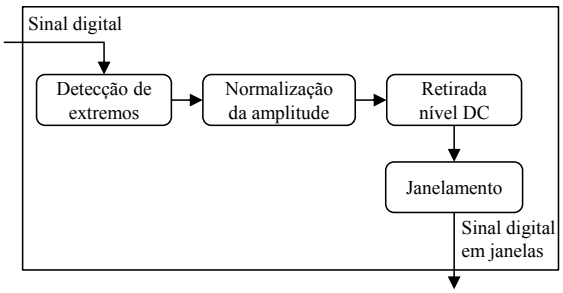
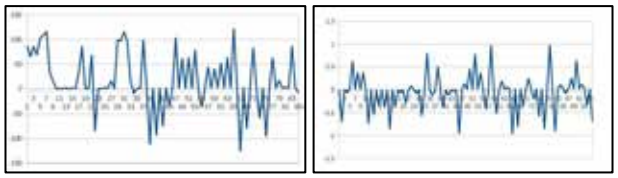


Figura 4: Operações no pré-processamento.

A *normalização* permite que sinais de amplitudes extremas, mais baixos e mais altos, tenham a mesma possibilidade de ser processadas que o resto. Na normalização todos os valores de amplitude do formato *wave* são transformados para valores flutuantes (*float*) na faixa de -1 e 1, dividindo o valor de cada amostra do sinal pelo maior valor de amplitude do mesmo. Na Figura 5 é apresentado o sinal digital da palavra falada "amarelo" e sua forma normalizada.



(a) (b)

Figura 5: Sinal da palavra "amarelo": (a) sinal digital após de detecção de extremos; (b) sinal digital normalizada.

Retirada de nível DC é um processo de eliminação da componente contínua (*DC*) do sinal que atrapalha a comparação em valores absolutos, de forma que todas as amostras fiquem oscilando em torno do valor 0. Esse processo é realizado subtraindo a cada amplitude a media aritmética das amplitudes do sinal digital.

O filtro de pré-ênfase realça as frequências do espectro de voz e melhora o desempenho da análise espectral para extração de características (Silva, 2009). É possível observar, em sinais de voz, que a energia presente nas frequências altas é menor se comparada às baixas frequências. Dado o vetor **x** de sinal digital deseja-se obter o vetor resultante da aplicação de pré-ênfase **y**, como $\mathbf{y}(n) = \mathbf{x}(n) - \alpha\mathbf{x}(n\text{-}1)$, sendo n índice da amostra e $\alpha = 0.95$, utilizados por Ruaro (2010) e Silva (2009).

No *janelamento*, o sinal resultante dos processos anteriores é dividido em intervalos de tempo mínimo para uma análise estacionária do sinal de fala. O janelamento vai permitir extrair um vetor de elementos descritores chamado *vetor de característica*, que neste trabalho denominamos como *vetor de referência*. Como o sinal é considerado quase estacionário em um intervalo de 10ms a 30ms (Rabiner, 1993), cada janela pode ser definida em intervalos menor que 30ms. Na literatura existem várias técnicas de janelamentos, como a retangular, Bartlett, Blackman, Hamming, e Welch. A janela de Hamming, utilizada neste trabalho, permite suavizar as bordas de cada segmento de intervalo quase estacionário. A aplicação da janela a um sinal do domínio de tempo corresponde à multiplicação do sinal pela função da janela representada. A função da janela é dada pela expressão, para N amostras da janela, como:

$$H(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{n-1}\right), & 0 \le n \le N-1 \\ 0, & n > N-1 \end{cases}$$

Essa função equivale a um filtro de forma de sino de Gauss que será aplicada sobre os intervalos de amostra da janela.

A sobreposição das janelas pode variar entre 0% a 70%. Quanto mais alta a sobreposição das janelas mais suave será transição dos parâmetros extraídos, porém, estimativas amplamente suavizadas podem ocultar variações reais do sinal e, caso a última janela ultrapassar os limites do sinal, deve-se complementar com zeros até final da janela. O tamanho utilizado neste trabalho é de 256 amostras por janela, com 50% de sobreposição. Na Figura 6 é ilustrada cinco janelas de Hamming complementando zeros até final da janela 5.

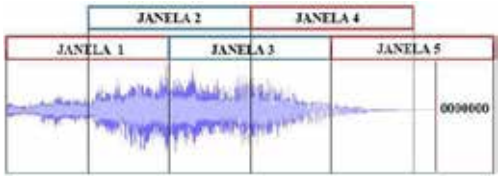


Figura 6: Janelas de Hamming com sobreposição de 50%.

3.3 Extração de características

Nesta etapa, trata-se de extrair um descritor representativo, com menor número de elementos, do sinal digital que contem muitas informações impossíveis de serem manipuladas nas operações de comparação, além disso, muitas informações existentes no sinal digital não possuem significância alguma para distinção fonética, assim o classificador empregado dificilmente conseguiria diferenciar amostras de palavras distintas.

Algumas técnicas de análise espectral são discutidas por Rabiner (1978), e são utilizadas para obter os parâmetros do sinal digital, elas são: a transformada rápida de Fourier (Fast Fourier Transform, FFT), os métodos de bancos de filtros (*Filter Bank*), os de análise homomórfica ou análise cepstral (mel-cepstrum), e os codificadores por predição linear (LPC: *Linear Predictive Coding*). As técnicas *FFT*, *Filter Bank* e *LPC* são utilizadas para a análise espectral da fala, no entanto, elas possuem algumas restrições, por isso Deller (1993) recomenda o uso da técnica mel-cepstrum, cujos coeficientes mel-cepstrais (MFCC: *Mel-Frequency Cepstral Coefficients*) são obtidos pela representação em frequência na escala Mel, a que considera a técnica mais apropriada para ser

utilizada no processo de reconhecimento de voz. Com vantagens no uso dessa técnica, atualmente os coeficientes MFCC são os mais populares (Bourouba, 2007).

Neste trabalho é utilizado o MFCC a partir das informações de transformada de Fourier aplicada em cada quadro de janelamento de Hamming. O processo seguido para gerar os coeficientes de MFCC, tal como ilustra a Figura 7, é o seguinte: análise espectral, divisão dos espectros em bandas utilizando banco de filtros Mel, logaritmo da energia de cada banda, e por último a transformada discreta de cossenos (DCT) na sequência de logaritmos para gerar os coeficientes de MFCC.

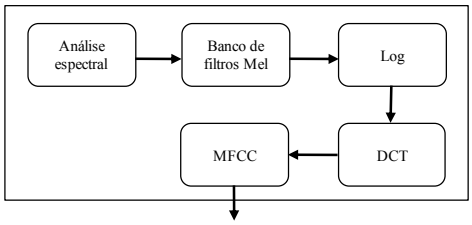


Figura 7: Extração de característica por MFCC.

Na *análise espectral*, a transformada discreta de Fourier (FDT) é aplicada a cada quadro resultante do janelamento de Hamming, obtendo-se o espectro de potências, que são os parâmetros mais úteis do sinal no domínio de frequências ao invés do domínio do tempo, permitindo uma distinção mais detalhada da composição fonética do sinal de som (SILVA, 2009). O espectro é dividido em bandas através dos filtros triangulares na escala Mel. Seguido a recomendação de Rabiner (1993), é utilizado 20 filtros no formato triangular passa-faixa, sendo 10 filtros uniformemente espaçados no eixo da frequência até 1000 Hz, e acima de 1000 Hz as faixas são distribuídas segundo uma escala logarítmica, como mostrada na Figura 8.

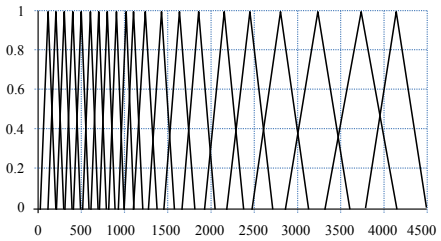


Figura 8: Banco de 20 filtros na escala Mel.

O valor do logaritmo da energia de cada banda relaciona o tipo de não linearidade observado no sistema auditivo humano. Por último, é aplicado DCT à sequência de logaritmos, a fim de descorrelatá-los, gerando os 12 coeficientes MFCC. Em sistemas de RAF normalmente são descartados alguns dos últimos coeficientes, pois isto provoca uma suavização do sinal. O primeiro coeficiente é função da soma das energias de todos os filtros, não tem importância não utilizar, então existiriam 11 coeficientes válidos.

4 Treino e reconhecimentos

Algumas amostras precisam ser geradas para serem utilizadas como padrões das palavras incluídas no dicionário e no modelo acústico. Também são realizados o treinamento e uma filtragem dessas amostras. Ince (1992) sugere dois tipos de padrão: Um tipo chamado *modelo estático* que faz um modelado estático das características exemplares do padrão, os HMMs são exemplos desse método. Outro tipo é conhecido como *padrão de referência não paramétrico*, podendo ser um exemplo do padrão a ser reconhecido ou um padrão médio do padrão a ser reconhecido.

4.1 Treino

Da forma como é treinado o sistema vai depender seu desempenho funcional. Para esse processo é definido um conjunto de falas, tendo por elementos a união de *n* falas de cada palavra do vocabulário, em seus respectivos coeficientes MFCC. Como o RAF proposto é independente de locutor, e de palavras isoladas, as falas são geradas de varias formas, como pronuncias rápidas, lentas, vozes diferentes, de homem e mulher, criança, jovem e adulto, etc. Inicialmente a coleta de dados de treino, *n* formas de cada cor, são gravados como arquivo de texto com nome da cor. Exemplo, “amarelo.txt”, “azul.txt”, “branco.txt”, “preto.txt”, “verde.txt”, e “vermelho.txt”.

O treino permite definir os padrões de referência do universo de informações tratadas, como neste caso, as falas relacionadas com a pronúncia das seis cores estabelecidas como referência do alfabeto do universo. Existem varias formas de representar esses padrões, neste caso conhecido como *conhecimento*, formas de representar estruturas com elementos ponderados a partir do conjunto de dados. Neste caso, como mencionado anteriormente, a representação é uma consequência das operações de treino a partir dos vetores característicos.

Para uso de modelo DTW na busca de um padrão desde a estrutura de representação de conhecimentos, é necessário estruturar os dados de treino em partições de classes no universo do domínio. Neste caso, particionar ou estruturar, todos os vetores de características das falas, em *K* partes. Para propósito neste trabalho, consideramos *K*=6, devido a que se deseja obter padrão representativo das seis cores. Utiliza-se o método K-Means para esse processo de partição. Para isso, K-Means requer referências iniciais, que geralmente obtêm aleatoriamente entre os dados de treino. Neste caso, por conveniência, é fornecida como entrada as referências iniciais com as melhores pronúncias das *K* falas, uma de cada cor, para forçar às iterações convergir em melhores referências das falas iniciais.

O método K-Means, amplamente utilizado por diferentes áreas no que diz respeito a agrupamentos, parte de *K* centros, neste caso referências, e agrupa todos os dados em *K* grupos em torno a cada centro, de forma que cada elemento (dado) de cada grupo tenha um grau de semelhança com o centro de referência do grupo. Terminado um processo de agrupamento, são recalculados novos centros em relação a cada grupo. Se existir alguma variação dos novos centros em relação aos centros utilizados para formar esses grupos, então são desfeitos os agrupamentos, e procede a uma nova agrupação de todos os elementos em torno aos novos

centros. Esse processo se repete, iterativamente, até convergir em centros mais representativos, ou seja, que não exista variação significativa dos novos centros em relação aos centros anteriores. O critério de comparação de semelhança é realizado pela função de distância, o seja a *menor distância* é interpretada como *semelhança*.

Como o método K-Means é busca local, cada padrão de referência encontrado para cada grupo vai estar próximo ao valor inicial forçada manualmente. Assim, a estrutura de dados de representação é um vetor de *K* elementos, cujos elementos são: o vetor que representa o padrão de referência, e a gramática de busca que vem a ser a cor desejada.

4.2 Classificação

A fala isolada, através do vetor de suas características, é comparada por similaridade em relação aos vetores característicos dos elementos padrão de referência definidos na etapa de treinamento. Se existir a similaridade, o padrão de referência mais próximo será considerado equivalente ao comparado; portanto, a palavra associada ao padrão de referência será ativada como a palavra reconhecida. O processo descrito é ilustrado pela Figura 9.

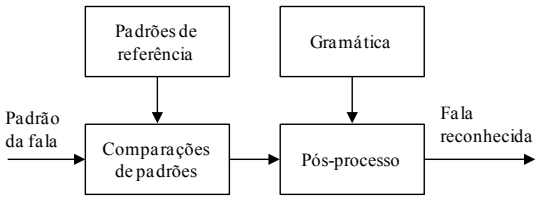


Figura 9: Reconhecimento baseado na comparação de padrões.

As comparações são realizadas pela função distância, tal como usado no processo de treino, conhecida também como DTW (em português, *deformação dinâmica no tempo*). A função distância, neste caso entre vetor de referência da fala a ser reconhecida e o vetor de referência padrão, pode ser a distância Euclidiana, distância Malahanobis, distância Bhattachryya, etc. Neste trabalho utilizamos a *distância Euclidiana* devido a sua simplicidade e baixo custo computacional que demanda. Assim, para um vetor de referência da fala $\boldsymbol{q}$ calculamos a distância para cada um dos *K* vetores referência padrão $\boldsymbol{p}_j$, a referência *r* escolhida é dada por

$$r = \text{argMin}\{\|\boldsymbol{q} - \boldsymbol{p}_j\|\}_{j=1,\dots,K}$$

Na realidade, no processo de implementação, é criado um vetor $\boldsymbol{d}$ com as *K* distâncias. Esse vetor $\boldsymbol{d}$ é ordenado em forma crescente, sendo, por tanto, $\boldsymbol{d}[0] \leq \boldsymbol{d}[1] \dots \leq \boldsymbol{d}[K-1]$, de forma que o primeiro elemento $\boldsymbol{d}[0]$ é a menor distância, o segundo elemento $\boldsymbol{d}[1]$ é segundo menor distância, etc. É lógico que cada posição de distância deve guardar o indicador *j* da referência padrão $\boldsymbol{p}_j$ ao qual corresponde a distância. Assim, $\boldsymbol{d}[0]$ é a distância para a referência padrão *r*.

Lembrar que essa referência é só um índice que indica que o padrão de referência $\boldsymbol{r}$ está mais próximo à referência da fala, mas não indica, necessariamente, a semelhança. Para verificar a semelhança relativa deve-se analisar o coeficiente de seletividade.

Os *coeficientes de seletividade* são utilizados para quantificar e qualificar uma determinada distância em relação às outras distâncias obtidas, e assim classificar se o reconhecimento pode ser considerado correto ou não. No reconhecimento a menor distância encontrada geralmente é a palavra reconhecida, mas palavras fora do banco de padrões podem ser ditas e seriam referenciadas a alguma menor distância, causando um erro. Por exemplo, se o usuário pronunciar a palavra "rosa", que não é considerada no banco de padrões, e a distância mínima encontrada indica ao padrão "azul", o resultado claramente é um erro.

Baseado no trabalho de Ruaro (2010), analisamos a qualidade de reconhecimento através dos coeficientes de seletividade com o objetivo de aceitar ou descartar o resultado. O processo de coeficiente de seletividade consiste em estabelecer dois coeficientes S_1 e S_2, da forma,

$$S_1 = \frac{\boldsymbol{d}[1]-\boldsymbol{d}[0]}{\boldsymbol{d}[0]} \quad \text{e} \quad S_2 = \frac{\bar{\boldsymbol{d}}-\boldsymbol{d}[0]}{\boldsymbol{d}[0]},$$

onde $\bar{\boldsymbol{d}}$ é a média das distâncias, $\boldsymbol{d}[0]$ e $\boldsymbol{d}[1]$ são as duas distância menores. Nesse caso, quando $\boldsymbol{d}[0]$ é menor e $\boldsymbol{d}[1]$ e $\bar{\boldsymbol{d}}$ forem maiores, melhor será a qualidade de reconhecimento. Neste trabalho, são estimados empiricamente, por testes sucessivos, os valores limites de aceitação, como $S_1 \geq 0.25$ e $S_2 \geq 0.08$.

5 Modelo implementado

A aquisição da fala é realizada com as funcionalidades nativas do sistema operacional *Android*, que possui algumas classes responsáveis por esse tipo de operação de dados e o microfone nativo do dispositivo móvel. A classe *AudioRecorder* é utilizada por ser mais flexível no processo de captura, e na edição dos valores de entrada. A fala é digitalizada e convertida em um arquivo de áudio *Pulse-Code Modulation* PCM no formato WAVE e salvo no cartão de memória do dispositivo. A digitalização é um processo de amostragem do sinal eletrônico da voz, como tal, expertos em tratamentos de som (Oliveira, 2002) recomendam deve ser entre 8KHz a 22KHz. Neste caso usou-se de 16KHz. Em consequência obtém-se um vetor de bytes com amplitudes relativas ao sinal sonoro, no formato codificado de 8 bits. O processo de *anti-aliasing* é realizado passando um filtro passa-baixa pela placa de som do dispositivo para minimizar as impurezas dos dados, assim as frequências mais altas passam a ser menores.

O formato de arquivo possui um cabeçalho que não possui utilidade na extração de características e se não for removido pode causar problemas no reconhecimento. Então é feita a remoção do cabeçalho do vetor retirando os índices de valor menor que 56 antes de iniciar a etapa de pré-processamento.

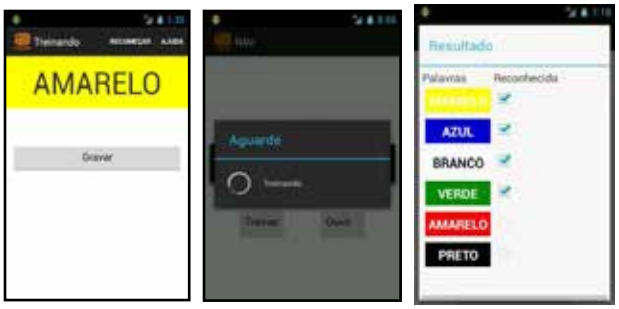
5.1 O aplicativo

O aplicativo foi testado em celulares Sansung Galaxy X, Samsung Galaxy4, tablet Motorola Xoom 2ME, no

trabalho de Ramos (2014). A Figura 10 ilustra a tela com o aplicativo. Lado esquerdo (a) é a primeira tela disponível para usuário com as funções básicas, onde "Jogar" permite ativar o processo de reconhecimento da palavra falada. O lado direito (b) mostra as cores do dicionário que podem ser reconhecidas.

Figura 10: Interface: (a) funções básicas; (b) palavras de cores permitidas no dicionário.

Para o processo de treino, é gravado um conjunto de falas das cores, como mencionado anteriormente, de varias formas, de 0 a 2 segundos de comprimento. Uma vez colecionado todas as falas de todas as cores, aciona-se "treinar". As palavras de falas iniciais para treino estão amarradas às palavras permitidas (Figura 10(b)). O que o processo de treino vai fazer é melhorar essas falas iniciais, de forma que sejam mais representativas de todas as pronúncias das cores correspondentes, sejam convertidas em referências padrão. Figura 11 mostra o treinamento, onde a parte (a) é processo de coleta das falas para treino, e lado (b) é propriamente o treino. Obviamente deve existir uma opção para remoção de pronúncias erradas que não foi implementada nesta versão.



(a) (b) (c)

Figura 11: Processo (a) coleta, treino (b), processo de acerto de falas de cores (c).

Como produto aplicativo, o objetivo é apenas reconhecer a palavra pronunciada, por isso a simplicidade do jogo. Quando o usuário clicar na opção **falar**, um tempo de 2 segundos é fornecido para pronuncia do nome da cor escolhida, o arquivo resultante dessa gravação é chamado de *elocução de teste*, passando pelo mesmo processo do treinamento com o pré-processamento e a extração de características o que se obtém é o vetor de referência, esse é comparado com as referências padrão, criados no treinamento, usando a distância euclidiana, método conhecido como DTW. Com as distâncias obtidas verifica-se a existência da distância válida pelo processo coeficiente de seletividade. Figura 11(c) ilustra o acerto das falas de cada cor.

5.2 Análise de casos de resultados

Em dispositivos móveis a variação do ambiente é um fator que dificulta o reconhecimento. Se o padrão é gravado em um ambiente silencioso e a elocução teste é dita em um ambiente ruidoso, o reconhecimento não é eficiente. Outra questão importante no reconhecimento é a questão do tempo de pronuncia, o sistema oferece 2 segundos para elocução das palavras, tempo mais que suficiente para pronunciar cada palavra isoladamente, o problema ocorre quando existe um retardamento na pronuncia da palavra. A palavra com atraso gera a falsa impressão para o usuário que ela foi capturada de forma correta, mas como o tempo de 2 segundos já havia se esgotado o sinal sonoro fica deformado em relação à amostra correta. Como consequência, nesse caso, a segunda fala dificilmente apontará amarelo.

As elocuções normalmente pronunciadas, após serem normalizadas, podem ser reconhecidas com certo grau de acertos. Por exemplo, falas de palavras cujas referências padrão ilustradas como na Figura 12(a, b, c), para amarelo, preto e vermelho, respectivamente; e testada para a fala vermelha com ruído (Figura 12(d)), o resultado é um reconhecimento correto, enquanto para a fala azul (Figura 12(e)) o resultado é errado.

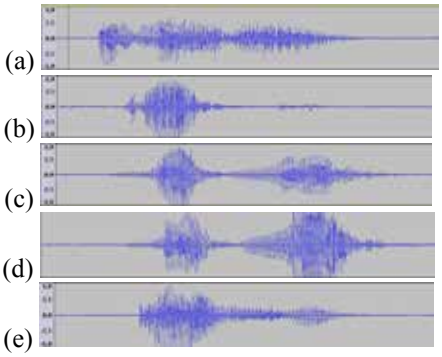


Figura 12: Espectros de elocuções de teste. Palavras padrão (a) amarelo, (b) preto, (c) vermelho; palavras teste (d) vermelho, (e) azul.

Na Tabela 1 são apresentados resultados da execução do sistema treinado por 4 pessoas: 2 do sexo feminino representadas por M1 e M2, e 2 do sexo masculino representados por H1 e H2.Cada locutor repetiu a elocução teste para cada palavra 5 vezes, obtendo-se melhores acertos variados, por H1 uma media de 80% em amarelo e vermelho e resto 60% Nos outros casos os acertos são menores, já que foram testados com falas diferentes do normal.

Tabela 1: Taxa de acertos nos testes.

	Amarelo	Azul	Branco	Preto	Verde	Vermelho
H1	80%	60%	60%	60%	60%	80%
M1	60%	40%	40%	40%	40%	60%
H2	40%	60%	40%	60%	40%	40%
M2	40%	40%	40%	60%	40%	60%

6 Conclusões e trabalhos futuros

O trabalho realizado tinha como objetivo a pesquisa e desenvolvimento de um sistema de reconhecimento de

voz, tendo como enfoque um sistema de reconhecimento de palavras isoladas.

As técnicas utilizadas na aquisição do sinal de voz, pré-processamento e extração de características deste projeto são as técnicas mais utilizadas na construção de sistemas de reconhecimento de voz. A técnica de comparações DTW se mostra eficiente apenas para um vocabulário pequeno, já que é feita uma comparação de distâncias entre cada padrão criado anteriormente e a elocução de teste, ou seja, um vocabulário grande geraria muitos padrões para serem comparados diminuindo a eficiência e desempenho do sistema.

O sistema de RAF implementado em Java, linguagem necessária para criação de aplicativos no sistema operacional Android, pode ser considerado dependente de locutor, já que um número considerável de amostras teria de ser criado no treinamento para possibilitar a utilização de qualquer locutor, diminuindo a eficiência do sistema e com reconhecimento para palavras isoladas pois o objetivo da aplicação é reconhecer comandos.

Como trabalhos futuros podemos considerar sistemas de reconhecedor independente do locutor, classificação com HMM, RNA e híbridos de forma a ampliar o dicionário; ajustes para incrementar a taxa de acertos. Também, poderia se *testar* combinações de valores na fase de aquisição de voz, como taxas de amostragens diferentes, número de canais, número e taxa de sobreposição das janelas, *modificar* o algoritmo de detecção de extremos para melhorar obtenção do sinal útil para reconhecimentos, e *ajustar novos parâmetros* para extração de características, como também usar outros métodos de extração.

Referências bibliográficas

Alvarenga, R. J. Reconhecimento de comandos de voz por redes neurais. Monografia em Ciência da Computação, Taubaté-SP, Brasil, 2-12.

Barcelos, A. Reconhecimento de voz para aplicação em cadeira de rodas. 2007. http://www.aedb.br/seget/artigos08/44 Reconhecimentodevozaplicadoemcadeiraderodas.pdf.

Bourouba, E.-H. Isolated words recognition system based on hybrid approach. MacMillan Publishing, 2007.

Bresolin, A. de A. Estudo do Reconhecimento de Voz para o Acionamento de Equipamentos Elétricos via Comandos em Português. Monografia (Mestrado), Joinville, Brasil, 2003.

Chu, W. C. Speech Coding Algorithms: Foundation and Evolution of Standardized Coders. [S.l.]: Wiley-Interscience, 2003.

Damasceno, E. F. Implementação de Serviços de Voz em Ambientes Virtuais, 2005. Disponível em: <http://www.dcc.ufla.br/infocomp/artigos/v4.3/art09.pdf>

Deller, J. R. Discrete-time processing of speech signals. Macmillan Publishing Company, 1993.

Ince, A. N. Digital speech processing: speech coding, synthesis, and recognition. Kluwer Academic Publishers, 1992.

Lee, Kai-Fu, Automatic Speech Recognition: The development of the Sphinx system. Klamer Academic Publishers, 1998, pags 211.

Louzada, J. A. Reconhecimento automático de fala por computador. Monografia em Ciências da Computação, PUC-Goiás, 2010.

Oliveira, K. M. Reconhecimento de voz através de reconhecimento de padrões.Monografia em Ciência da Computação, Salvador-BA, Brasil, 2012.

Paula, M. B. Reconhecimento de palavras faladas utilizando Redes Neurais Artificiais. Monografia em Ciência da Computação, Pelotas, 2000.

Rabiner, L. R. Fundamentals of speech recognition. Ed. Prentice Hall, 1993.

Rabiner, L. R. Digital processing of speech signals. Ed. Prentice Hall, 1978.

Ramiro, P. H. de O. Sistema de acionamento de dispositivos comandado por voz. Monografia em Ciência da Computação, Brasília, 2010.

Ramos, Raphael. Reconhecimento de fala isolada em dispositivos móveis, monografia de Bacharel em Ciências da Computação, Universidade Estadual do Norte Fluminense – UENF, Brasil, 2014.

Rodrigues, F. F. Acionamento de um robô lego por comandos vocais utilizando redes neurais artificiais. Monografia em Ciência da Computação, Ouro Preto, Brasil, 2009.

Ruaro, M. Obtenção da frequência de um sinal de som por meio da fft em java me. Santo Ángelo – RS, 2010.

Saha, G.; Chakroborty, S.; Senapati, S. A new silence removal and endpoint detection algorithm for speech and speaker recognition applications. India, 2005.

Sena, M. Nacimento, A. Comparação de Técnicas de Classificação Utilizando a Distância de Mahalanobis Amostral com Técnicas de Detecção de Outliers, 19 SINAPE, 26-30 de julho, São Pedro, São Paulo, Brasil, 2010.

Silva, A. G. Reconhecimento de voz para palavras isoladas. Monografia em Ciência da Computação, Recife, 2009.

Zue, Victor. The lese of speech knowledge in automatic speech recognition. Proceedings of the IEEE, v. 73, n. 11, 1985, pags 1602-1615.