Técnicas de Extração de Características da Fala

Leonardo Manhães Gomes

Instituto de Computação – Universidade Federal Fluminense (UFF)

Resumo

Este documento apresenta as seguintes técnicas de extração de características da fala: MFCC, PLP e LPC. Estas técnicas são usadas principalmente para reconhecimento automático da fala ou do orador. Discute-se também o emprego de outras técnicas onde são relatados estudos que realizam comparações de acurácia entre tais técnicas de extração. Além disso, são comentados trabalhos que descrevem possíveis soluções com emprego do reconhecimento da fala.

Keywords: técnicas de extração de características da fala, reconhecimento da fala, reconhecimento de orador.

1. Introdução

A fala é um dos principais meios de comunicação natural humano. A tecnologia possibilitou a criação de meios de comunicação artificiais que agregaram em nossa evolução tecnológica e possibilitaram potencializar e dinamizar a comunicação verbal. Em qualquer protocolo ou pacote de dados que está sendo transmitido por uma rede de computadores há informações que precisam ser recebidas, interpretadas e reconhecidas ou entendidas no destino para que haja uma transmissão correta da informação. De maneira similar, para que haja uma boa comunicação entre duas pessoas, quando uma pessoa fala, a outra precisa receber, interpretar e entender a informação verbal pronunciada. Assim como uma informação dentro de um pacote de redes é formada por bits, as informações verbais contêm características próprias da fala que permitem o reconhecimento da palavra, do sentimento e da pessoa que está falando. Para reconhecer esta individualidade de palavras, sentimentos ou pessoas, existem técnicas de extração de características da fala que permitem trabalhar tais características da fala de um orador e apresentá-las a um sistema ou algoritmo em um formato transformado, mais fácil de ser processado. Tais técnicas são utilizadas em sistemas que podem apresentar uma diversidade de objetivos relacionados ao reconhecimento da fala, de palavras, idiomas, sentimentos e oradores.

O objetivo deste trabalho é registrar os benefícios da utilização das técnicas de extração de características de fala, em especial a técnica MFCC, quando executadas em conjunto com métodos de classificação na criação de sistemas que tem como objetivo o reconhecimento de fala, idioma, orador ou até mesmo de emoções.

Este documento apresenta na Seção 2 algumas técnicas de extração de característica da fala. Na Seção 3, são referenciados artigos que realizaram pesquisas comparando as performances de tais técnicas de extração. A Seção 4 realiza referências aos artigos que trazem possibilidades de soluções distintas, no âmbito do reconhecimento da fala com

uso das técnicas de extração de características da fala. Tais soluções podem ser futuramente representas na forma de sistemas multimídia. Na Seção 5, são registradas as conclusões e ideias para trabalhos futuros.

2. Técnicas de Extração de Características da Fala

O primeiro passo em qualquer técnica de extração de características da fala é extrair e identificar os componentes do sinal de áudio que são adequados para identificar o conteúdo linguístico e descartar todas as outras partes que transportam informações desnecessárias como por exemplo ruído de fundo.

O ponto principal a ser entendido sobre a fala é que os sons gerados por um humano são filtrados pela forma do trato vocal, incluindo a língua, os dentes, etc. Essa forma determina o som que sai. Se pudermos determinar a forma com precisão, isso deve nos dar uma representação precisa do fonema que está sendo produzido. A forma do trato vocal se manifesta no envelope do curto espectro de tempo e o trabalho das técnicas de extração é representar com precisão este envelope.

Esta seção apresenta a técnica de extração MFCC e mais outras duas técnicas de extração.

2.1 - MFCC

No processamento de som, o cepstrum de frequência mel (*mel-frequency cepstrum* - MFC) é uma representação do espectro de potência de curto prazo de um som. O MFC foi criado por Bridle e Brown [Bridle e Brown 1974] onde eles usaram um conjunto de 19 coeficientes em formato de espectro ponderados dados pela transformada de cosseno das saídas de um conjunto de filtros de passagem de banda não uniformemente espaçados. A diferença entre o cepstrum e o cepstrum de frequência mel é que no MFC as bandas de frequência são igualmente espaçadas na escala mel, tornando esta representação mais próxima do comportamento do sistema auditivo humano do que as bandas de frequência linearmente utilizadas no cepstrum normal.

Como o sinal de fala consiste em tons com frequências diferentes, para cada tom com uma frequência real f, medida em Hz, um tom subjetivo é medido na escala mel.

A escala mel, nomeada por Stevens, Volkmann e Newman em 1937 [Stevens et al. 1937] é uma escala perceptual da altura do som, ou o tom (característica do som que varia de grave a agudo), inspirada no sistema auditivo humano. O ponto de referência entre esta escala e a medição de frequência habitual é definido atribuindo um tom perceptual de 1.000 mels a um tom de 1.000 Hz, 40 dB acima do limiar do ouvinte [Hassan et al. 2004]. O nome mel vem da palavra melodia.

Os coeficientes cepstrais de frequência mel (MFCCs), que são coeficientes que coletivamente formam uma MFC, foram então introduzidos por Davis e Mermelstein na década de 1980 [Davis e Mermelstein 1980]. Têm como objetivo a captura de características do som e são amplamente utilizados em projetos para reconhecimento automático de fala e de oradores.

O uso de cerca de 20 coeficientes MFCC é comum em ASR (*Automatic Speech Recognition*), embora entre 10 a 12 sejam frequentemente considerados suficientes para codificar a fala [Hagen et al. 2003].

A técnica de computação do MFCC é baseada na análise de curto prazo e assim que cada quadro (frame) do áudio é processado, um vetor MFCC é dada como saída deste processo disponibilizando os coeficientes cepstrais do áudiocomputado. A fim de extrair os coeficientes, a amostra de fala é tomada como a entrada e a janela de hamming é aplicada para minimizar as descontinuidades de um sinal. Então a Transformada Discreta de Fourier (*Discrete Fourier Transform* – DFT) será usada para gerar o banco de filtros Mel. O MFCC pode ser calculado usando a fórmula apresentada na Equação (1) para converter a frequência em Hertz para a correspondente escala Mel [Kumar et al. 2014].

$$Mel(f) = 2595 * \log10(1 + f / 700) \quad (1)$$

O passo a passo para a geração do MFCC se dá da seguinte forma:

PASSO 1: PRÉ-ÊNFASE

No pré-processamento, componentes de alta frequência são enfatizados. Este processo aumentará a energia de sinal em frequência mais alta, conforme a Equação (2). [Kakade e Salunke 2018]

$$y(n) = x(n) - \alpha * x(n-1) \quad (2)$$

Onde x (n) é o sinal de fala de entrada e $0,9 \leq \alpha \leq 1$. Normalmente $\alpha = 0,95$, o que faz com que 95% de qualquer amostra seja presumida como originária da amostra anterior.

PASSO 2: Enquadramento (Framing)

O enquadramento envolve a segmentação de amostras de fala obtidas após a digitalização dos sinais de fala em um pequeno quadro com o comprimento dentro do intervalo de 20 a 40 ms. O sinal de voz é dividido em quadros de N amostras. Quadros adjacentes estão sendo separados por M ($M < N$). Valores típicos utilizados são $M = 100$ e $N = 256$ [Kakade e Salunke 2018].

PASSO 3: Janelamento de Hamming

Entre vários tipos de janelas, a janela de Hamming é usada considerando o próximo bloco na cadeia de processamento de extração de características e integra todas as linhas de frequência mais próximas. A janela de Hamming é calculada como descrito a seguir [Kakade e Salunke 2018].

Se a janela é definida como $w(n)$, $0 \leq n \leq n-1$, onde:

n = número de amostras em cada quadro

y(n) = sinal de saída

x(n) = sinal de entrada

w(n) = janela de Hamming, então o resultado do sinal de janelamento é mostrado na Equação (4):

$$y(n) = x(n) * w(n) \quad (3)$$

$$w(n) = 0,54 - 0,46 * \cos(2 \pi n / n-1); \text{ onde } 0 < n < n-1 \quad (4)$$

PASSO 4: Transformada Rápida de Fourier (*Fast Fourier Transform* – FFT)

Nesta etapa, a Transformada Rápida de Fourier de cada quadro no domínio de tempo é convertida para a representação no domínio da frequência. A magnitude do sinal de janela é calculada para obter o espectro de potência [Dhonde e Jagade 2015].

PASSO 5: Conversão para o Espectro na Escala Mel

Nesta etapa, o sinal de janela é multiplicado por uma escala de frequências não linear chamada escala Mel, que é aproximadamente linear até 1 kHz e logarítmica acima de 1 kHz. A relação entre a frequência linear e a escala mel é dada pela Equação (5) [Dhonde e Jagade 2015],

$$Mel(f) = 2595 * \log10(1 + f / 700) \quad (5)$$

PASSO 6: Transformada Discreta de Cosseno (*Discrete Cosine Transform* – DCT)

O DCT é usado para converter o espectro do log Mel para o domínio do tempo. O resultado da conversão é chamado Coeficiente Cepstrum de Frequência Mel. O conjunto de coeficiente é chamado de vetores acústicos. Portanto, cada enunciado de entrada é transformado em uma sequência de vetores acústicos. O DCT é preferido em relação ao IFFT (Inverse Fast Fourier Transform), porque para uso do IFFT é necessário um grande número de cálculos, o que aumenta a complexidade [Kakade e Salunke 2018].

A representação cepstral do espectro de fala fornece uma boa representação das propriedades espectrais locais do sinal para a análise de quadro (frame) de fala. Como os coeficientes do espectro mel são números reais (e assim são seus logaritmos), eles podem ser convertidos para o domínio do tempo usando a Transformada Discreta de Cosseno (DCT). O número de coeficientes mel cepstrum é tipicamente escolhido como 20. O primeiro componente é excluído do DCT, uma vez que representa o valor médio do sinal de entrada que transporta pouca informação específica do falante. Aplicando o procedimento descrito acima, para cada quadro de fala de cerca de 30 ms com sobreposição, é calculado um conjunto de coeficientes de cepstrum de frequência mel. Esses vetores acústicos podem ser usados para representar e reconhecer a característica de voz do locutor [Hassan et al. 2004].

Antes da introdução de MFCCs, Coeficientes de Predição Linear (LPCs) e Coeficientes Cepstrais de Predição Linear (LPCCs) eram o principal tipo de recurso para sistemas de Reconhecimento Automático de Fala (*Automatic Speech Recognition* - ASR).

2.2 Outras técnicas de extração

2.2.1 Codificação Preditiva Linear (*Linear Preditive Coding* – LPC)

A predição linear é uma operação computacional matemática que é uma combinação linear de várias amostras anteriores. O LPC da fala tornou-se a técnica predominante para estimar os parâmetros básicos da fala. Ele fornece uma estimativa precisa dos parâmetros de fala e também é um modelo computacional eficiente da fala. A ideia básica por trás do LPC é que uma amostra de fala pode ser aproximada como uma combinação linear de amostras de fala passadas. Por meio da minimização da soma das diferenças quadradas (em um intervalo finito) entre as amostras reais de fala e os valores previstos, um conjunto exclusivo de parâmetros ou coeficientes preditores pode ser determinado. Esses coeficientes formam a base para o LPC da fala [Kumar et al. 2014]. A Figura 1 apresenta o esboço de execução do processo LPC.

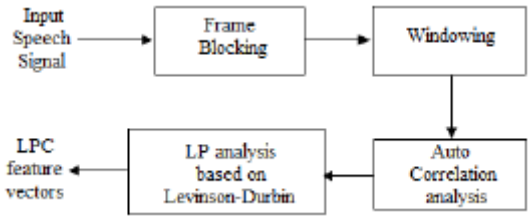

Figura 1 – Diagrama de bloco da Codificação Preditiva Linear [Kumar et al. 2014]

2.2.2. Predição Linear Perceptual (*Perceptual Linear Prediction* - PLP)

O modelo de PLP foi desenvolvido por Hermansky em 1990. Ele modela o discurso humano baseado no conceito de psicofísica da audição. O PLP descarta informações irrelevantes da fala e melhora assim a taxa de reconhecimento de voz. O PLP é idêntico ao LPC, exceto que suas características espectrais foram transformadas para corresponder às características do sistema auditivo humano [Dave 2013].

Ele é uma representação popular no reconhecimento de fala e baseia-se no espectro de curto prazo da fala. Parâmetros de PLP são os coeficientes que resultam da modelagem padrão de todos os polos, que é eficaz na supressão de detalhes específicos do orador do espectro. Além disso, a sequência PLP é menor do que a normalmente é necessário para sistemas de reconhecimento de fala baseados em LPC. No PLP, o espectro de fala é modificado por um conjunto de transformações baseadas no modelo do sistema auditivo humano. As etapas de cálculo do PLP são: a resolução espectral de banda crítica, a curva de audição de volume igual e a lei da potência da intensidade sonora da audição. Uma vez que o espectro do tipo auditivo é estimado ele é convertido em valores de autocorrelação através da Transformada de Fourier. As autocorrelações resultantes são usadas como entrada para uma rotina de análise preditiva linear padrão e sua saída são coeficientes de predição linear baseados na percepção. Normalmente, esses coeficientes são então convertidos em coeficientes cepstrais por meio de uma recursão padrão [Singh et al. 2014]. A Figura 2 apresenta o esboço de execução do processo PLP.

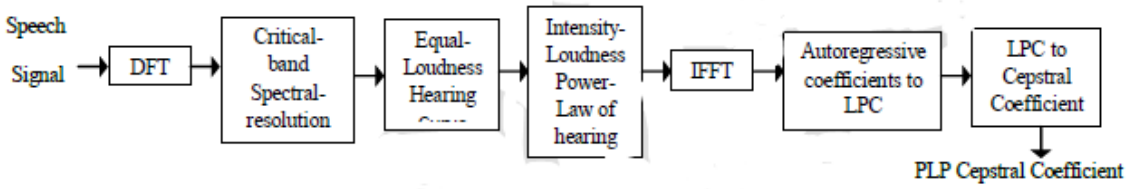

Figura 2 - Implementação PLP [Singh et al. 2014]

3. Comparativo entre MFCC e outras técnicas de medições

Nesta seção, serão mostrados resultados de comparações entre técnicas de extrações de características de fala realizadas por estudos anteriores, tendo como objetivo mostrar a performance de resultados nas contribuições destas técnicas para o reconhecimento de fala ou de orador.

Mehta L. et al. (2013) um estudo comparativo entre as técnicas MFCC e LPC para o reconhecimento de palavras isoladas da língua Marathi (língua indo-ariana, falada na

Índia ocidental e central). O banco de dados de fala Marathi é gravado em ambiente ruidoso, visando a aplicação de uma ferramenta de aprendizado de idiomas. O banco de dados consiste em palavras marathi simples que começam com vogais e consoantes. Cada palavra foi repetida 10 vezes por um orador masculino e um feminino. Para a identificação dos oradores foi utilizado um método de quantização vetorial baseado na distância euclidiana. As Tabelas 1 e 2 apresentam os comparativos dos resultados no reconhecimento das palavras utilizando MFCC e LPC.

Tabela 1: Acurácia de reconhecimento para características usando LPC [Mehta et al. 2013]

WORD	SPEAKER 1	SPEAKER 2
AAI	75%	73%
ANANAS	78%	74%
BAL	80%	78%
KSHATRIYA	81%	80%
AVERAGE	78.5%	76.25%

Tabela 2: Acurácia de reconhecimento para características usando MFCC [Mehta et al. 2013]

WORD	SPEAKER 1	SPEAKER 2
AAI	98%	99%
ANANAS	100%	100%
BAL	100%	100%
KSHATRIYA	100%	100%
AVERAGE	99.5%	99.75%

Outro estudo para reconhecimento de palavras indianas foi realizado por Gulzar T. et al. (2014). Foi realizada uma análise comparativa das técnicas de extração de características de fala MFCC, LPCC e BFCC para o reconhecimento de palavras Hindi (língua indo-ariana falada principalmente na Índia central e norte). Para o processamento do reconhecimento das palavras foi utilizada uma rede neural artificial. As Figuras 3 e 4 exibem gráficos registrando o percentual de acurácia de cada técnica para o reconhecimento das palavras.

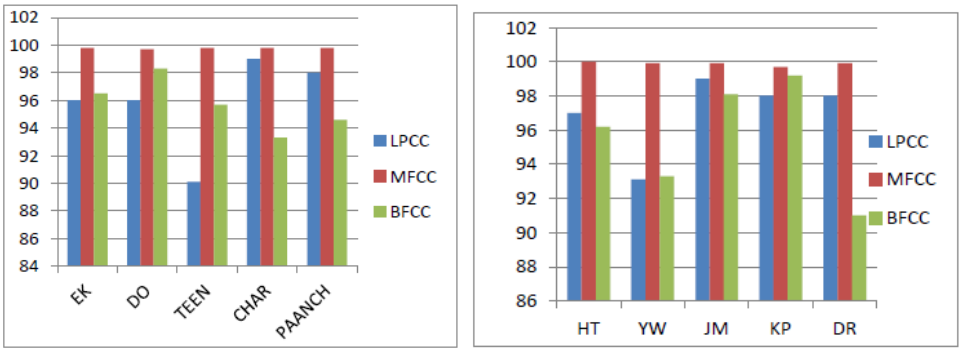

Figura 3 – Acurácia para o reconhecimento de palavas: à esquerda reconhecimento de palavras isoladas; à direita reconhecimento de pares de palavras (representadas no gráfico apenas pela primeira letra) [Gulzar et al. 2014]

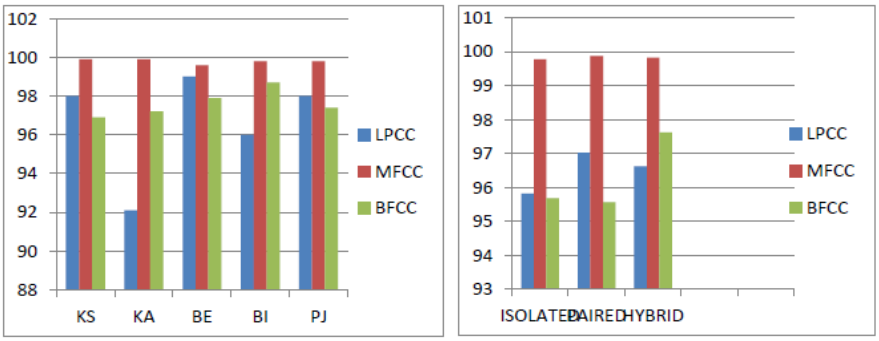

Figura 4 – Acurácia para o reconhecimento de palavras: à esquerda reconhecimento de palavras híbridas (formação aleatória de pares de palavras, representadas apenas pela primeira letra); à direita a média dos 3 gráficos anteriores [Gulzar et al. 2014]

Kumar J. et al. (2014) realizaram uma análise comparativa utilizando diferentes técnicas de extração de características de fala e métodos classificadores para identificação de oradores. As técnicas de extração comparadas foram: MFCC, LPC e PLP. Os métodos classificadores comparados foram: Redes Neurais Artificiais, Modelo Oculto de Markov, Modelo Oculto de Markov Híbrido, Distância Euclidiana, DTW (*Dynamic Time Warping*). A comparação é exibida na Figura 5.

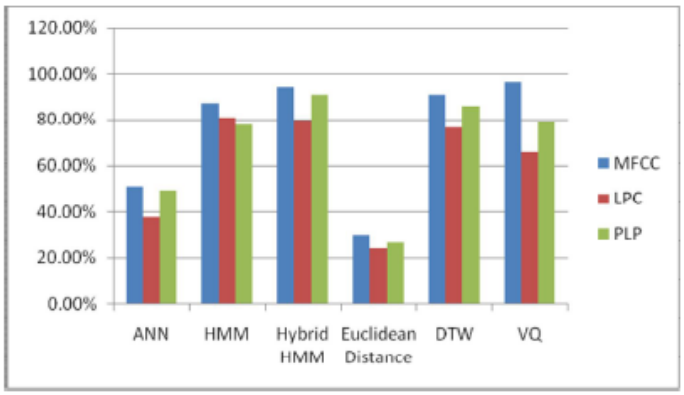

Figura 5 - Taxa de reconhecimento utilizando diferentes técnicas de extração de características e classificadores [Kumar et al. 2014]

Uma análise comparativa da utilização individual e combinada de técnicas de extração de características da fala foi feita por Hasan R. et al. em [Hassan et al. 2017]. A base de dados consiste em 2000 falas de quatro palavras árabes isoladas pronunciadas por 50 oradores árabes nativos sendo que cada orador repete a palavra 10 vezes. As técnicas de extração utilizadas são MFCC, LPCC e PLP e a combinação entre elas, resultando em 6 tipos de extrações. A Figura 6 exibe o gráfico mostrando a taxa de acurácia no reconhecimento das falas.

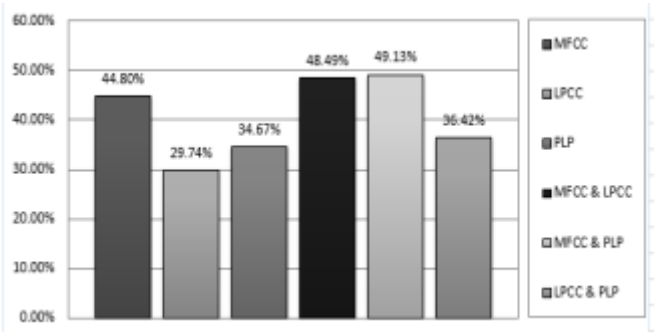

Figura 6 – Taxa de reconhecimento para aplicação das técnicas individualmente ou combinadas [Hassan et al. 2017]

4. Sistemas Multimídia que utilizam MFCC

A contribuição das técnicas de extração de características da fala nas pesquisas para reconhecimento de fala e de oradores foi imensa. A evolução no ramo de reconhecimento de fala e de orador abre um leque de possibilidades para a criação de novos sistemas, incluindo sistemas multimídia, explorando várias possibilidades de inovação. Esta seção apresenta alguns exemplos de estudos que concretizam estas possibilidades.

4.1 - Uma nova abordagem para detecção de cópia de vídeo usando impressões digitais de áudio e PCA

Roopalakshmi e Reedy (2011) preocuparam-se com o grande aumento de publicações de vídeos na Internet e o desafio crítico que é realizar o controle de direitos autorais desta imensa quantidade de vídeos. Eles destacam que na literatura de Detecção de Cópia Baseada em Conteúdo (*Content-Based Copy Detection* - CBCD), várias técnicas de controle já foram criadas, mas a imensa maioria delas baseadas na mídia vídeo. Exemplos são recursos baseados em palavras visuais, análise comparativa de histogramas de cores, métodos baseados em borda para detectar cópias de vídeo, estudos comparativos com utilização de vários descritores globais e locais.

Visto isso, o trabalho [Roopalakshmi e Reedy 2011] propõe a criação de um novo método para controle de direitos autorias. O algoritmo proposto inclui dois passos: primeiro é realizada a extração da impressão digital do áudio da mídia utilizando a técnica MFCC para a obtenção das características do áudio. Durante este processamento, o sinal de áudio é reduzido à taxa de amostragem de 22050 Hz, a fim de reduzir o tamanho dos dados a serem processados. Logo depois quatro descritores espectrais são calculados para capturar características baseadas no áudio: descritor de distribuição espectral, energia do sinal, Roll-off espectral e Fluxo espectral.

No segundo passo, vários vetores de recursos são processados usando o PCA (*Principal Component Analysis*) para fornecer uma representação de recurso compacta. Os resultados da detecção demonstram a eficiência do método proposto em relação a diferentes edições e transformações de vídeo.

4.2 – Classificação musical usando MFCC e SVM

No trabalho [Thiruvengatanadhan 2018], Thiruvengatanadhan propõe um método para classificação musical automática considerando a extração das características da música através da técnica MFCC e classificando a música com SVM (*Support Vector Machine*).

Segundo o autor esta solução traria melhorias nos serviços de indexação de músicas, recuperação de músicas baseada em conteúdo, recomendação musical e distribuição de música on-line. Além da melhoria nestes serviços, a preocupação do autor está também na organização automática dos repositórios de músicas, automatizando o trabalho manual de classificação musical, categorizando informações de música por diferentes critérios, como artista, gênero, sub-gênero e similaridade musical. Além disso, o próprio gênero poderia estar sendo atualizado, por exemplo, na categoria rock deverão estar armazenados rocks dos anos 60 e rocks atuais, embora saibamos que os rocks atuais apresentam estilos bem diferentes do que eram tocados nos anos 60.

Os resultados experimentais mostraram que o método proposto de aprendizado de SVM para áudio tem bom desempenho no esquema de classificação de gênero musical apresentando uma taxa de precisão de 93%.

4.3 - Abordagem sobre Reconhecimento Automático de Fala para Fonoaudiologia de Pacientes Afásicos

Jamal N. et al. (2017) examinam o desenvolvimento recente de tecnologias ASRs e seu desempenho para indivíduos com distúrbios de fala e linguagem mantendo um foco em pacientes afásicos. A afasia é uma condição na qual a pessoa afetada sofre de distúrbio de fala e linguagem resultante de um derrame ou lesão cerebral. Com a evolução de sistemas ASR surgiram várias soluções e possibilidades para contribuição em diversos ramos inclusive comunicação, medicina, terapia e fonoaudiologia. O ASR é uma tecnologia que tem como uma de suas possibilidades transferir a fala humana para texto transcrito. Isso é particularmente útil em terapias de reabilitação da fala, uma vez que fornecem uma avaliação precisa e em tempo real para a fala de um indivíduo com distúrbio da fala. As abordagens baseadas em ASR para terapia fonoaudiológica reconhecem a entrada de fala do paciente afásico, fornecem resposta de feedback em tempo real aos seus erros e pode praticar exercícios cognitivos. No entanto, a precisão da ASR depende de muitos fatores, tais como, reconhecimento de fonemas, continuidade de fala, diferenças ambientais e de oradores, bem como o profundo conhecimento sobre a compreensão da linguagem humana.

A Figura 7 mostra a taxa de erro encontrada em trabalhos anteriores sobre abordagem baseada em ASR, na utilização de técnicas de extração de características da fala em conjunto com modelos para reconhecimento do dado acústico.

Authors	Speech Task	Feature Extraction	Acoustic Modelling	Error Rate
Le *et. al.* [51]	Continuous	MFCC-LDA	GMM-HMM	39.7%
			DNN-HMM	42.9%
Lee *et. al.* [6]	Continuous	MFCC-LDA	GMM-HMM	58.2%
			DNN-HMM	57.8%
Abad *et. al.* [4]	Isolated	PLP-Rasta-Modulation Spectrogram (MSG)	MLP-HMM	21.0%

Figura 7 – Taxa de erros de trabalhos anteriores sobre abordagem baseada em ASR para pacientes com afasia

4.4 - Reconhecimento de Emoção na Fala Usando Parâmetros de *Fourier*

Levando em consideração várias abordagens de pesquisas relacionadas ao reconhecimento de emoções na fala, Wang et al. (2015) ainda veem como necessárias mais contribuições de estudo na interpretação de emoções presentes na voz e propõem uma nova técnica. Levando em consideração a teoria musical, onde a estrutura da harmonia de um intervalo ou acorde é principalmente responsável por produzir uma impressão positiva ou negativa nos ouvintes, esse artigo propôs um conjunto de sequências harmônicas, chamadas de parâmetros de Fourier (FP), para detectar o conteúdo perceptivo das características de qualidade de voz, em vez do conteúdo convencional. Os novos recursos do FP são avaliados em diferentes bancos de dados de fala. Foram utilizadas a classificação bayesiana e a SVM (*Support Vector Machine*) para a avaliação.

Com isso, as principais contribuições desse artigo para reconhecimento de emoções do orador são:

1) propor um novo modelo de FP usando recursos de FP e suas diferenças de uma e segunda ordem para o reconhecimento de emoções de fala;

2) propor a melhorar ainda maior do reconhecimento de emoções de fala, independente do orador, combinando recursos FP e MFCC;

3) realizar extensas validações em três bases de dados de fala em dois idiomas: banco de dados alemão (EMODB), banco de dados chinês (CASIA) e banco de dados de idosos chineses (EESDB).

Avaliando emoções de felicidade (*happiness*), tristeza (*sadness*), zanga (*angry*), descontentamento (*disgust*), medo (*fear*), neutralidade (*neutral*) e surpresa (*surprise*) e utilizando técnicas de extração de características da fala (MFCC, FP e combinações de mais de uma técnica), foram obtidos os resultados exibidos na Figura 8 como taxas de reconhecimento das emoções sobre os 3 bancos de dados.

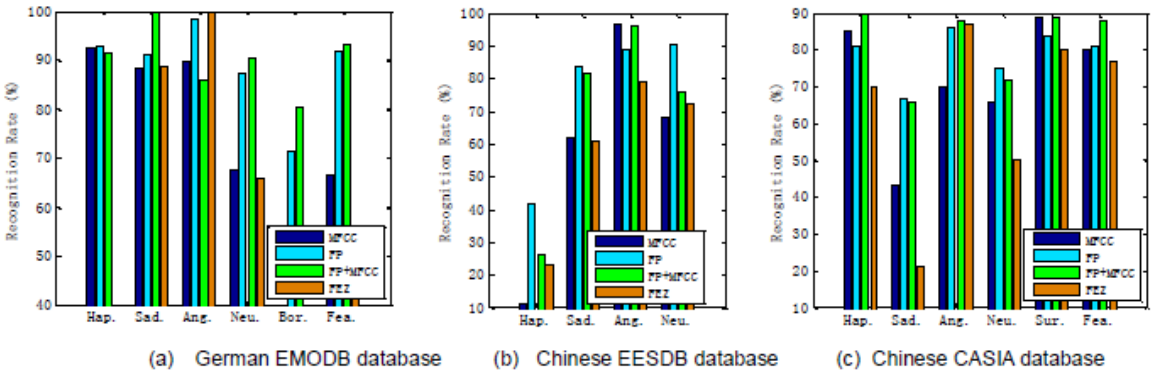

Figura 8 – Taxas de reconhecimento de emoções nos idiomas Alemão e Chinês utilizando técnicas de extração de características de fala.

Segundo Wang et al.(2015), o estudo mostrou que as características da FP são eficazes na caracterização e reconhecimento de emoções nos sinais de fala independente do orador.

5. Conclusão e trabalhos futuros

Com o que foi apresentado neste trabalho, pôde-se notar a importância que as técnicas de extração de características da fala têm para os algoritmos de reconhecimento de fala e de

orador. Pôde-se conhecer um pouco sobre algumas das técnicas de extração e ver resultados com empregos de tais técnicas demonstrados por outros trabalhos. Além disso, foram apresentados artigos que mencionavam soluções para emprego de reconhecimento de fala. As possibilidades de criação de sistemas future11dia com o processamento do reconhecimento de fala são imensas.

Como trabalho future, há a possibilidade de desenvolvimento de pesquisas para comparação do emprego das técnicas de extração de características da fala demonstradas neste documento e de outras existentes para o reconhecimento de oradores.

Referências

[1] Bridle J. S., Brown M. D. (1974), "An Experimental Automatic Word-Recognition System", JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, England.

[2] Stevens S. S., Volkmann; J., Newman, E. B. (1937). "A scale for the measurement of the psychological magnitude pitch". Journal of the Acoustical Society of America. 8, 185–190 (1937).

[3] Hasan R., Jamil M., Rabbani G., Rahman S. (2004), "SPEAKER IDENTIFICATION USING MEL FREQUENCY CEPSTRAL COEFFICIENTS", 3rd International Conference on Electrical & Computer Engineering 2004, 28-30 December 2004, Dhaka, Bangladesh.

[4] Davis S. B., Mermelstein P. (1980), "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," in IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), pág. 357–366.

[5] Hagen A., Connors D.A., Pellm B.L. (2003), "The Analysis and Design of Architecture Systems for Speech Recognition on Modern Handheld-Computing Devices". Proceedings of the 1st IEEE/ACM/IFIP international conference on hardware/software design and system synthesis, pág. 65-70.

[6] Kumar J., Prabhakar O., Sahu N. (2014), "Comparative Analysis of Different Feature Extraction and Classifier Techniques for Speaker Identification Systems: A Review", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1, January 2014.

[7] Kakade M., Salunke D. (2018), "Real Time Speaker Independent Speech Recognition System", International Journal of Innovations & Advancement in Computer Science (IJIACS) - Volume 7, Issue 3 - March 2018.

[8] Dhonde S., Jagade S. (2015), "Feature Extraction Techniques in Speaker Recognition: A Review", International Journal on Recent Technologies in Mechanical and Electrical Engineering (IJRMEE) ISSN: 2349-7947 Volume: 2 Issue: 5 – pág. 104 a 106.

[9] Mehta L., Mahajan S., Dabhade A. (2013), "COMPARATIVE STUDY OF MFCC AND LPC FOR MARATHI ISOLATED WORD RECOGNITION SYSTEM", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering - Vol. 2, Issue 6, June 2013.

[10] Gulzar T., Singh A., Sharma S. (2014), "Comparative Analysis of LPCC, MFCC and BFCC for the Recognition of Hindi Words using Artificial Neural Networks", International Journal of Computer Applications (0975 – 8887) Volume 101– No.12, September 2014.[11] Hasan R., Hussein H., Lazaridis P. et al. (2017), "Improvement of Speech Recognition Results by a Combination of Systems", Proceedings of the 23rd

International Conference on Automation & Computing, University of Huddersfield, Huddersfield, UK, 7-8 September 2017.

[12] Roopalakshmi R., Reddy G. (2011), "A Novel Approach to Video Copy Detection Using Audio Fingerprints and PCA", The 2nd International Conference on Ambient Systems, Networks and Technologies (ANT-2011), Procedia Computer Science 5 (2011) 149–156.

[13] Thiruvengatanadhan R. (2018), "Music Classification using MFCC and SVM", International Research Journal of Engineering and Technology (IRJET) - Volume: 05 - Issue: 09 - September 2018.

[14] Jamal N., Shanta S., Mahmud F., Sha'abani M. (2017), "Automatic Speech Recognition (ASR) based Approach for Speech Therapy of Aphasic Patients: A Review", AIP Conference Proceedings - Published by the American Institute of Physics.

[15] Wang K., An N., Li B., Zhang Y., Li L. (2015), "Speech Emotion Recognition Using Fourier Parameters", IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, 6(1):69–75, Jan.

[16] Dave N. (2013), "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition", INTERNATIONAL JOURNAL FOR ADVANCE RESEARCH IN ENGINEERING AND TECHNOLOGY, Volume 1, Issue VI, July 2013.

[17] Singh V., Jain V., Tripathi N. (2014), "A Comparative Study on Feature Extraction Techniques for Language Identification", International Journal of Engineering Research and General Science Volume 2, Issue 3, April-May 2014