Classificação de Sotaques em Áudio

José Geraldo Fernandes
Escola de Engenharia
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
josegeraldof@ufmg.br

***Resumo*—Este trabalho propõe uma forma de classificação de sotaques a partir do áudio. Dois modelos de classificação foram testados e a extração de características é pela média em *frames* de coeficientes do espectro em frequência mel (MFCC). Dois conjuntos de dados foram utilizados, o *Common Voice - English* e o Corpus CEFALA-1. Todo o código desenvolvido está disponível em repositório Git.**

***Palavras-chaves*—classificação de sotaques, mfcc, mlp, anfis**

I. Introdução

Com a integração de sistemas inteligentes na vida social e o esforço de atendimento personalizado surge um problema de representatividade grave [1], [2]. A maioria de sistemas baseados em aprendizado sofre de um viés pesado originário dos dados. A diferenciação de acesso econômico e social implicou uma discriminação nas formas de coleta de dados. A partir disso, é comum desses sistemas ter um desempenho inferior para classes pouco representadas.

A partir disso, há um incentivo de aumento dessa coleta de dados, senão técnicas para eliminar esse viés do conjunto de dados. E, juntamente, uma capacidade desses sistemas de identificar agrupamentos específicos e agir de acordo.

Por isso, é cada vez mais popular uma técnica de classificação de sotaques a partir de áudio [3]. Em uma aplicação de assistente virtual, por exemplo, regionalismos mal representados podem impactar o desempenho do sistema, principalmente quando há diferenças conflitantes.

Este trabalho apresenta uma abordagem de classificação de regionalismos nacionais, e internacionais como teste de hipótese, unicamente a partir do áudio.

A Seção II mostra os conjuntos de dados utilizados, a extração de características dos áudios e os modelos de classificação. A Seção III apresenta os resultados, de forma geral e entre classes. E, as Seções IV e V apresentam as discussões e conclusões do trabalho.

II. Metodologia

A. Base de Dados

Para classificação dos sotaques duas bases de dados foram selecionadas. Como teste de hipótese, a *Common Voice - English* [4] foi selecionada por ser um conjunto massivo e discretizar *labels* de sotaque na forma de países falantes da língua inglesa. Essa base de dados é composta por contribuições públicas e mantém seu acesso livre. Selecionou-se os cinco sotaques mais predominantes do *dataset* e em proporção original para simular uma situação realista. A Tabela I mostra o número de amostras selecionadas para cada classe.

Tabela I
Proporção de amostras selecionadas por classe do *Common Voice - English*.

Classe	Amostras
us	600
england	400
indian	250
canada	150
australia	100

Para validação, utilizou-se a Corpus CEFALA-1 [5], uma base de dados em português com 104 locutores. As amostras são áudios em formato espontâneo e roteirizado dos locutores. Para classificação dos sotaques, fez-se a avaliação manual de todos os locutores. Por conta da origem, a maior parte dos falantes foi classificado como mineiro. A Tabela II mostra a relação de locutores e amostras por sotaque.

Tabela II
Proporção de falantes e amostras por classe do Corpus CEFALA-1.

Classe	Falantes	Amostras
mineiro	84	1680
baiano	9	180
paulista	7	140
estrangeiro	2	40
nortista	1	20
nordestino	1	20

B. Extração de Características

Como o formato bruto dos áudios é pouco tratável por métodos de classificação rasos, extraiu-se os coeficientes do espectro em frequência mel (MFCC) como em [3].

Para normalizar a representação do sinal em função da frequência aplicou-se um filtro pré-ênfase, a relação sinal-ruído típico da voz depende da frequência, e uma transformação para escala mel, em função da percepção acústica humana. Os áudios são separados em *frames* e uma transformada de Fourier de tempo curto é aplicada em cada. Finalmente, os atributos do problema são a média dos primeiros $q = 39$ coeficientes.

C. Classificação

O problema de classificação segue o padronizado. Para cada *dataset* separou-se 10 *folds* fixos, desses aplicou-se dois modelos de classificação, uma *Adaptive-Network-Based Fuzzy Inference System* (ANFIS) [6] e uma *Multi-Layer Perceptron* (MLP) [7]. O problema foi tratado como uma classe contra todas, no caso do Corpus CEFALA-1 apenas a classe majoritária testada, por conta da maioria desproporcional.

Avaliou-se a performance com a área sob a curva ROC (AUC) com validação cruzada nos *folds*. Também, discriminou-se essa métrica por classe separando o *subset* que contém o par específico.

III. Resultados

As Tabelas III e IV mostram os resultados obtidos com a AUC média dos *folds* para os sotaques das bases de dados.

Tabela III
DESEMPENHO OBTIDO NA BASE DE DADOS *Common Voice - English*.

Sotaque	**MLP**	**ANFIS**
us	0.606	0.615
england	0.647	0.628
indian	0.736	0.782
canada	0.609	0.611
australia	0.654	0.664

Tabela IV
DESEMPENHO OBTIDO NA BASE DE DADOS CORPUS CEFALA-1.

Sotaque	**MLP**	**ANFIS**
mineiro	0.845	0.902

Também avaliou-se a AUC média por par de classes no *Common Voice - English*. O resultado é como nas Tabelas V e VI.

Tabela V
DESEMPENHO DE CLASSES CRUZADOS DA MLP.

	us	england	indian	canada	australia
us	-	0.611	0.708	0.688	0.700
england	0.553	-	0.655	0.571	0.678
indian	0.786	0.764	-	0.708	0.886
canada	0.595	0.590	0.633	-	0.623
australia	0.744	0.722	0.732	0.742	-

Tabela VI
DESEMPENHO DE CLASSES CRUZADOS DA ANFIS.

	us	england	indian	canada	australia
us	-	0.645	0.701	0.683	0.676
england	0.622	-	0.745	0.659	0.519
indian	0.777	0.753	-	0.766	0.903
canada	0.605	0.591	0.654	-	0.643
australia	0.775	0.772	0.689	0.726	-

IV. Discussões

Mostrou-se que os atributos MFCC modelam bem o problema de classificação de sotaques e classificadores simples como o MLP e ANFIS alcançaram um resultado descente.

Houve um diferencial grande entre o desempenho dos dois *datasets*, há algumas hipóteses para explicar esse fenômeno. A caracterização do problema é diferente, espera-se que a discriminação de regionalismos internacionais seja mais simples que sotaques nacionais, esse é um fator contrário à evidência, primeiro ponto. Apesar disso, segundo ponto, a qualidade de áudio do Corpus CEFALA-1 é muito superior. Isso é razoável já que enquanto o *Common Voice* parte da coleta em *crowdsourcing*, o conjunto de dados em português utilizou-se diversos equipamentos especializados para a aquisição. Por fim, e terceiro ponto, as amostras de texto roteirizado podem ser vantajosas para a classificação que o formato espontâneo, embora isso seja uma desvantagem em um sistema de classificação prático.

Outro evento interessante é o desempenho entre classes no *Common Voice - English*. Observou-se, na prática, uma medida de semelhança entre os sotaques em função de quão misturadas estão as amostras e, em consequência, quão bem enganam o modelo de classificação. O sotaque indiano que passa a impressão de ser mais único, de fato mostra essa intuição nos resultados, por exemplo.

V. Conclusões

Este trabalho apresentou uma tentativa de discriminação de sotaques a partir unicamente de áudio do falante. A extração de características se mostrou decente para o trabalho, apesar que propriedades da aquisição de dados impactaram significativamente o desempenho.

Uma continuação deste trabalho é natural em três dimensões: a classificação; o conjunto de dados; e, a extração de características. Na primeira, o teste com outros modelos de classificação é simples e o aumento do desempenho é possível. Na segunda, conjunto de dados em português se mostraram escassos em comparação aos internacionais, um esforço para coleta e até classificação manual de outros conjuntos seria benéfica para o problema; finalmente, outras técnicas de extração de características poderiam ser empregadas em vez dos atributos MFCC, como por exemplo *Linear Predictive Coding* (LPC) [8] e análise *formant* [9], ou até mesmo trabalhar com todos os coeficientes em vez da média dos *frames*.

Referências

[1] A. Schlesinger, K. P. O'Hara, and A. S. Taylor, "Let's talk about race: Identity, chatbots, and ai," in *Proceedings of the 2018 chi conference on human factors in computing systems*, pp. 1–14, 2018.

[2] C. L. Bennett, C. Gleason, M. K. Scheuerman, J. P. Bigham, A. Guo, and A. To, ""it's complicated": Negotiating accessibility and (mis) representation in image descriptions of race, gender, and disability," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2021.

[3] Z. Ma and E. Fokoué, "A comparison of classifiers in performing speaker accent recognition using mfccs," *arXiv preprint arXiv:1501.07866*, 2015.

[4] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[5] A. F. Neto, A. P. Silva, and H. C. Yehia, "Corpus cefala-1: base de dados audiovisual de locutores para estudos de biometria, fonética e fonologia/corpus cefala-1: audiovisual database of speakers for biometric, phonetic and phonology studies," *Revista de Estudos da Linguagem*, vol. 27, no. 1, pp. 191–212, 2019.

[6] J.-S. Jang, "Anfis: adaptive-network-based fuzzy inference system," *IEEE transactions on systems, man, and cybernetics*, vol. 23, no. 3, pp. 665–685, 1993.

[7] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[8] D. O'Shaughnessy, "Linear predictive coding," *IEEE potentials*, vol. 7, no. 1, pp. 29–32, 1988.

[9] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *The Journal of the Acoustical Society of America*, vol. 47, no. 2B, pp. 634–648, 1970.