Determinação da valência e classificação dos sons emitidos por bebês utilizando pyAudioAnalysis

Wagner Machado do Amaral
Departamento de Engenharia de Computação e Automação Industrial (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, SP, Brasil

INTRODUÇÃO

Recém-nascidos emitem sons de maneira diferente de acordo suas necessidades. Conforme Lindova [5], humanos são capazes de distinguir com precisão se os sons emitidos pelos bebês correspondem a situações positivas ou negativas. Com menor precisão é feita a estimativa da situação específica, possivelmente com a combinação de pistas visuais e acústicas, como expressões faciais, intensidade e valência do som.

Algoritmos de detecção automática e classificação dos sons emitidos pelos bebês podem auxiliar os pais a entender e atender às necessidades dos bebês. Algoritmos de detecção automática de um choro infantil, como o proposto por Cohen [2], podem ser utilizados para a identificação de um perigo físico para os bebês, como situações em que os pais deixam seus filhos em veículos. Outros algoritmos podem ainda distinguir entre diferentes tipos de choros de bebês e estimar a possível causa, como fome, eructação, cansaço ou dor [1][3][4][6][7].

O presente trabalho propõe um algoritmo para determinar a valência do som emitido por um bebê como positiva ou negativa. Em caso de valência negativa, o algoritmo classifica a necessidade do bebê como: fome, eructação, dor, sono, frio/calor ou desconforto (possivelmente causado por solidão ou medo).

OBJETIVO

Desenvolver um algoritmo para determinar a valência do som emitido por um bebê como positiva ou negativa. Em caso de valência negativa, o algoritmo classifica a necessidade do bebê como: fome, eructação, dor, sono, frio/calor ou desconforto.

DESENVOLVIMENTO

Neste trabalho foram utilizados dois repositórios de sons produzidos por bebês para treinar o algoritmo e para validar o resultado. O primeiro [9] foi construído através da campanha Donate-a-cry e contém amostras de áudio, em sua forma original, carregadas pelo usuário usando o aplicativo móvel Donate-a-cry para Android e iOS. Os áudios estão organizados de acordo com dados informados pelos usuários, como o gênero do bebê, idade e motivo do choro. O segundo repositório [10] contém uma coleção de vídeos de situações que envolvem bebês ou crianças rindo.

O desenvolvimento do algoritmo é dividido em 4 etapas, conforme é exibido na Figura 1. Inicialmente, é realizada uma etapa de sanitização da base de dados, removendo os áudios considerados inadequados ao contexto do trabalho e selecionando os arquivos para serem utilizados nas etapas seguintes. Em seguida, na etapa de pré-processamento, os dados que estão em formato de vídeos são manualmente convertidos em áudio e os trechos contendo sons de bebês são recortados. Parte da base então é utilizada para treinar um algoritmo construído em Python utilizando a biblioteca pyAudioAnalysis [11]. Finalmente, o algoritmo é aplicado à base toda e os resultados são avaliados.

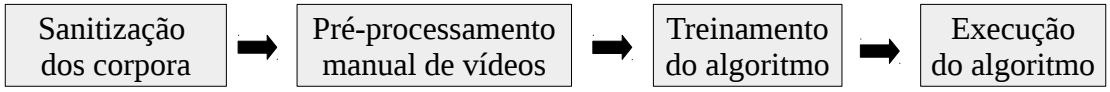

Figura 1: Fluxograma do desenvolvimento do trabalho

Na primeira etapa exibida na imagem 1, todos os 1128 áudios da base Donate-a-cry corpus [9] foram avaliados manualmente. Destes, 775 foram considerados inadequados e foram descartados pelas seguintes razões:

- Áudio não continha som de bebê;
- Áudio continha som de adulto imitando bebê;
- Áudio continha som de criança com mais de 2 anos;
- Áudio classificado pelo usuário como “Don’t know”, que corresponde a situação na qual o usuário não sabe a razão do choro do bebê.

Após a eliminação dos áudios inadequados, os 353 arquivos restantes foram classificados como: Fome, Eructação, Dor, Sono, Frio/Calor e Desconforto. As cinco primeiras classes foram herdadas do corpus, e estão relacionadas a sensações físicas conhecidas, como necessidade de alimentação, necessidade de arrotar, dores de barriga, cansaço e incomodo causado pela temperatura. A última, chamada Desconforto, engloba os áudio do corpus classificados como desconforto, solidão e medo. Estas três classes do corpus foram unidas por não terem uma razão física evidente do choro.

Para compor os sons de bebês rindo, 6 vídeos foram extraídos da base Baby Laughter AudioSet [10]. Os vídeos foram manualmente processados utilizando a ferramenta Audacity [12] e as partes correspondentes a risos de bebês foram transformadas em arquivos de áudio, totalizando 49 sons de bebês rindo. A tabela 1 exibe informações sobre os corpora de origem e a base de dados utilizada. No restante do trabalho, as classes são referenciadas pelos seus respectivos rótulos, compostos por 2 caracteres associados ao nome da classe em inglês (la, hu, bu, bp, ti, ch e dc).

Corpus de origem	**Rótulo no corpus de origem**	**Rótulo utilizado no presente trabalho**		**Quantidade**
		Rótulo	**Significado**	
Baby Laughter AudioSet [10]	Baby Laughter	la	Riso	49
Donate-a-cry corpus [9]	Hungry	hu	Fome	277
	Needs burping	bu	Eructação	5
	Belly pain	bp	Dor	15
	Tired	ti	Sono	22
	Cold/Hot	ch	Frio/Calor	7
	Discomfort	dc	Desconforto	27
	Lonely			
	Scared			
	Don't know	---	---	---
			Total	402

Tabela 1: Composição do corpus utilizado

RESULTADOS

Para treinar o algoritmo foram selecionados 10 áudios de cada classe, exceto para as classes bu e ch, que totalizam menos de 10 áudios. Conforme é exibido na tabela 2, a base de treinamento foi formada por um conjunto de 62 áudios de classes distintas.

Cartegoria	**Quantidade**
bp	10
bu	5
ch	7
dc	10
hu	10
la	10
ti	10
Total	62

Tabela 2: Tamanho da base de treinamento

Após o término do treinamento do algoritmo, todos os 402 áudios foram avaliados e classificados. A tabela 3 exibe a correlação entre as classes reais e as classes estimadas. A diagonal da matriz corresponde aos casos de acerto e expressa, para cada classe, o total de áudios que receberam do algoritmo uma classe igual à sua respectiva classificação real na base de dados. Os demais valores da matriz correspondem aos falsos positivos. Os falsos negativos podem ser calculados a partir da subtração do total de áudios de uma classe na base de dados pela quantidade de acertos de estimativa para a mesma classe. Por exemplo, para os 15 áudios correspondentes à classe “bp”, apenas 4 foram classificados corretamente pelo algoritmo.

		Estimado							
		bp	**bu**	**ch**	**dc**	**hu**	**la**	**ti**	**Total**
Real	**bp**	4	0	0	6	2	0	3	**15**
	bu	4	0	0	1	0	0	0	**5**
	ch	1	0	5	1	0	0	0	**7**
	dc	13	0	1	6	7	0	0	**27**
	hu	109	1	14	58	61	5	29	**277**
	la	0	0	0	0	0	49	0	**49**
	ti	9	0	1	5	4	1	2	**22**
	Total	**140**	**1**	**21**	**77**	**74**	**55**	**34**	**402**

Tabela 3: Correlação entre quantidade real e estimada por classificação

A partir dos dados exibidos na tabela 3 é possível determinar a taxa de precisão da determinação da valência dos áudios. Para esse fim, considera-se que a valência foi positiva quando o algoritmo associou o áudio à classe “la” (riso) e negativa quando o áudio foi associado às demais classes (choro). Conforme é exibido na tabela 4, a determinação da valência foi correta para 98,51%

do áudios. Os erros correspondem a 6 casos de falso positivo ocorridos em vídeos que continham ruídos ou vozes de adultos, o que confundiu o algoritmo.

Além de determinar a valência, que é considerada positiva em caso do áudio ser considerado um riso e negativa em caso de ser um choro, o algoritmo estimou a causa do choro. Conforme é exibido na tabela 4, o algoritmo acertou a classificação dos diferentes tipos de choros em 22,10% dos áudios com valência negativa.

Precisão da determinação da valência	Precisão da classificação dos sons negativos
98,51%	22,10%

Tabela 4: Taxas de precisão da determinação da valência e classificação

A baixa taxa de acerto obtida na classificação dos tipos de choro é atribuída ao tamanho pequeno da base de treinamento e a possíveis contaminações nos dados, visto que os áudios correspondentes a choros foram doados por usuários que fizeram a classificação de forma intuitiva.

A Figura 2 exibe, para cada classe, o número real de áudios, o número de áudios corretamente classificados pelo algoritmo e o número de falsos positivos. Além da instabilidade das taxas de acerto para a classificação dos tipos de choro, é possível observar a diferença de amostras dos corpora. Uma base de dados maior e com amostras distribuídas de forma mais homogênea entre as classes pode contribuir para a melhoria do resultado final.

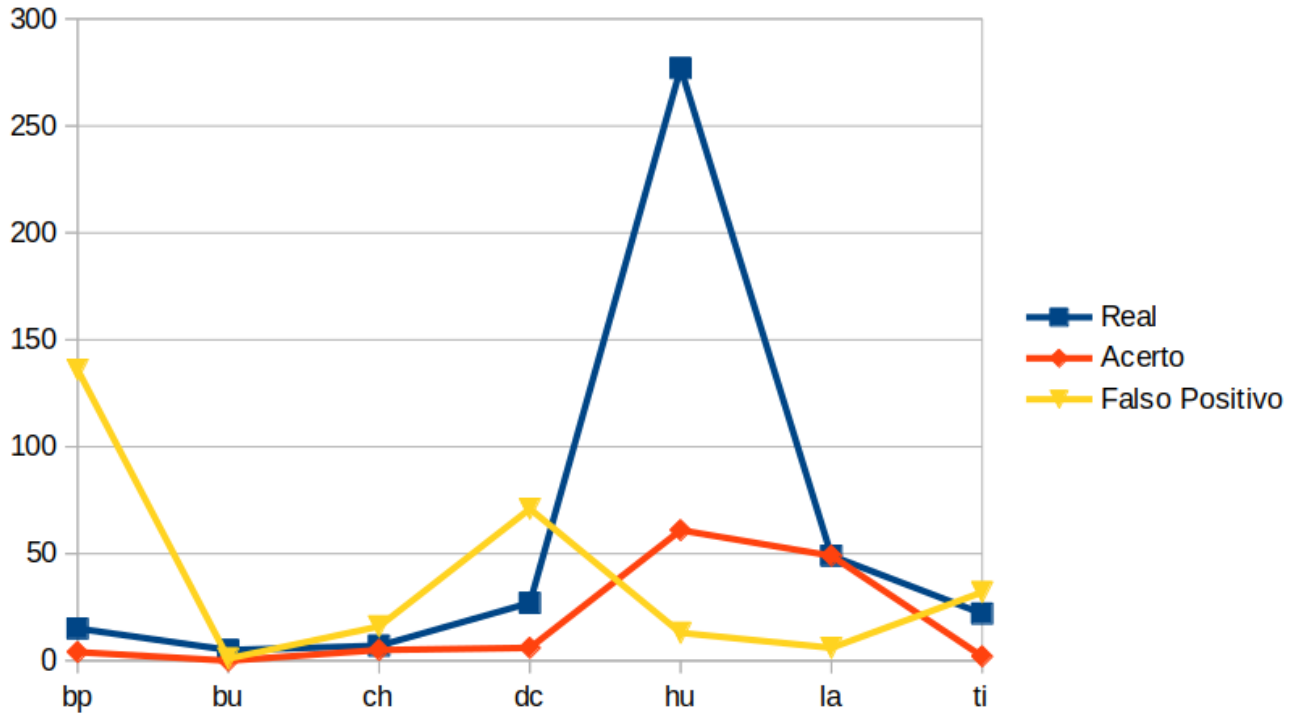

Figura 2: Relação entre total de áudios e resultado da classificação por classe

CONCLUSÃO

O presente trabalho propôs um algoritmo para determinar a valência do som emitido por um bebê como positiva ou negativa. Em caso de valência negativa, o algoritmo classifica a necessidade do bebê como: fome, eructação, dor, sono, frio/calor ou desconforto. Resultado promissores foram obtidos na diferenciação entre choro e riso, que determina a valência do som. Resultados mais precisos podem ser obtidos na classificação dos tipos de choro utilizando bases de treinamento maiores e de melhor qualidade.

Possíveis melhorias incluem a utilização de técnicas automáticas de remoção de ruídos, detecção automática de trechos do áudio correspondentes a sons de bebês e a portabilidade do algoritmo para dispositivos móveis, o que facilitaria o uso da ferramente em situações cotidianas.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Banica, Ioana-Alina & Cucu, Horia & Buzo, Andi & Burileanu, Dragos & Burileanu, Corneliu. (2016). Automatic methods for infant cry classification. 51-54. 10.1109/ICComm.2016.7528261.

[2] Cohen, Rami & Lavner, Yizhar. (2012). Infant cry analysis and detection. IEEE 27th Convention Oj Electrical and Electronics Engineers in Israel. 1-5. 10.1109/EEEI.2012.6376996.

[3] Jagtap , Sandhya & Kadbe , Premanand & Arotale, Parshuram. (2016). System Propose For Be Acquainted With newborn Cry Emotion Using Linear Frequency Cepstral Coefficient. ICEEOT.2016.

[4] Kikuchi, K & Arakawa, Kaoru. (2005). Estimation of babies' emotion by frequency analyses of their cries. 7-. 10.1109/NSIP.2005.1502217.

[5] Lindova, Jitka & Spinka, Marek & Martinec Novakova, Lenka. (2015). Decoding of Baby Calls: Can Adult Humans Identify the Eliciting Situation from Emotional Vocalizations of Preverbal Infants?. PLoS ONE. 10. e0124317. 10.1371/journal.pone.0124317.

[6] Tejaswini, S & Sriraam, Natarajan & C M Pradeep, G. (2016). Recognition of infant cries using wavelet derived mel frequency feature with SVM classification. 1-4. 10.1109/CIMCA.2016.8053313.

[7] Yamamoto, Shota & Yoshitomi, Yasunari & Tabuse, Masayoshi & Kushida, Kou & As, Taro. (2013). Recognition of a Baby's Emotional Cry Towards Robotics Baby Caregiver. International Journal of Advanced Robotic Systems. 10. 1. 10.5772/55406.

[8] Yamamoto, Shota & Yoshitomi, Yasunari & Tabuse, Masayoshi & Kushida, Kou & Asada, Taro. (2010). Detection of baby voice and its application using speech recognition system and fundamental frequency analysis. 341-345.

[9] Donate-a-cry Infant cry audio corpus.
https://github.com/gveres/donateacry-corpus

[10] Baby Laughter AudioSet.
https://research.google.com/audioset/eval/baby_laughter.html

[11] Giannakopoulos T (2015) pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. PLoS ONE10(12): e0144610.

[12] http://www.audacityteam.org/