O Problema do Reconhecimento de Voz *Offline* em Dispositivos Móveis: em Busca de uma Abordagem Racional

Lucas Debatin
Universidade do Vale do Itajaí
Itajaí - SC - Brasil
lucasdebatin@edu.univali.br

Aluizio Haendchen Filho
Centro Universitário de Brusque
Brusque - SC - Brasil
aluizio.h.filho@gmail.com

Rudimar L. S. Dazzi
Universidade do Vale do Itajaí
Itajaí - SC - Brasil
rudimar@univali.br

ABSTRACT

This paper presents a systematic review of the literature on the subject of continuous offline voice recognition for Android mobile devices. We examined 222 articles from four digital repositories, which followed a methodology containing search questions, search expression, and inclusion and exclusion criteria. After reading the abstract, introduction and conclusion, 12 articles were selected and synthesized through answers to the research questions. Based on the synthesis, we provide information on how to select the best practices for neural networks utilization, and also suggest techniques for reducing error rate.

KEYWORDS

Continuous speech recognition, offline voice recognition, mobile

1 INTRODUÇÃO

O reconhecimento de voz contínuo é o mais complexo e difícil de ser implementado, pois deve ser capaz de lidar com todas as características e vícios de linguagem da forma natural [1]. Os avanços nas técnicas de reconhecimento de voz viabilizam o uso desta tecnologia em diversas aplicações, sobretudo em dispositivos móveis. Os portadores de necessidades especiais também são beneficiados com tais sistemas, usuários que não podem usar as mãos e deficientes visuais usam essa tecnologia para se expressarem, realizando o controle sobre várias funções do computador por meio da voz.

No mercado atual, há diversas API's (*Application Programming Interface*) que facilitam a implementação do reconhecimento de voz em *softwares* e aplicações, por exemplo, Web *speech*, Java *speech*, Google *cloud speech*, dentre outras. Entretanto, nenhuma delas realizam o reconhecimento em modo *offline*, ou seja, é necessário que o usuário esteja conectado à *internet*. Uma limitação é que muitas dessas API's são *softwares* proprietários, ou seja, não são gratuitas, e em muitos casos o valor pago se torna alto, pois depende diretamente da quantidade de requisições que a API realiza. O reconhecimento de voz também está sendo muito utilizado em *softwares* de empresas, e, em muitos casos, há a necessidade de que o mesmo seja *offline* e gratuito.

Este trabalho apresenta uma revisão sistemática da literatura em busca de uma abordagem racional para o desenvolvimento de uma solução de reconhecimento *offline* de voz aplicada a dispositivos móveis com Android. Uma abordagem racional destaca os benefícios do produto e suas características, apontando suas qualidades, desempenho e economia.

2 METODOLOGIA

Os artigos obtidos foram selecionados em função dos seguintes critérios: (*i*) Quais redes neurais, dentro desta área de reconhecimento de voz, estão sendo mais pesquisadas; (*ii*) Quais soluções estão sendo estudadas para reduzir as taxas de erros de reconhecimento de voz; (*iii*) É utilizado o modelo estatístico *n-gram* para aperfeiçoar o reconhecimento de voz; e (*iv*) Quais maneiras para disponibilizar de modo *offline* o reconhecimento de voz em dispositivos móveis com Android.

Para responder as estas perguntas, foram selecionados quatro repositórios de pesquisa: (i) IEEE; (ii) ACM; (iii) Scopus; e (iv) ScienceDirect. Utilizou-se a seguinte expressão de busca: (("*speech recognition*" **OR** "*offline speech recognition*" **OR** "*continuous speech recognition*" **OR** ((*mobile* **OR** android) **AND** "*speech recognition*")) **AND** ("*neural network*" **OR** "*deep learning*") **AND** ("*n-gram*" OR ("*natural language processing*" **OR** nlp))).

Ao analisar esta expressão, é possível observar que foram utilizadas palavras chaves que remetem ao problema do estudo. Em seguida, foram definidos os critérios de escolha dos artigos:

1. Critérios de inclusão: (*i*) artigos publicados entre 01/01/2012 até 30/06/2017; (*ii*) expressão de busca filtrando os artigos através do título, resumo e palavras chaves;
2. Critérios de exclusão: (*i*) artigos curtos (resumos expandidos); (*ii*) conferências específicas sem corpo revisor público; (*iii*) artigos que possuem, em mais de um ano de publicação, menos de dez citações; (*iv*) artigos que não apresentam o uso de alguma rede neural; e (*v*) artigos em idiomas diferentes do inglês e do português.

Após realizou-se a seleção desses artigos com a leitura dos seguintes tópicos: (*i*) título e palavras chaves utilizadas; (*ii*) resumo; e (*iii*) introdução e conclusão. Esses critérios para a leitura e seleção foram úteis para minimizar o esforço na leitura e

In: Sessão de Pôsteres do WebMedia´17, Gramado, Brasil. Anais do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web: Workshops e Pôsteres. Porto Alegre: Sociedade Brasileira de Computação, 2017.

ISBN: 978-85-7669-380-2.

seleção de trabalhos que realmente contribuem para responder às perguntas do tema de pesquisa.

A Tabela 1 apresenta os artigos selecionados, cada qual com a sua referência (utilizada para identificar os artigos), número de citações, ano, e as três primeiras perguntas de pesquisa que, por meio da leitura de cada artigo, foram respondidas de maneira sintetizada, visando facilitar a interpretação e a tabulação dos dados. A quarta pergunta, relacionada com a utilização de reconhecimento *offline* não aparece na tabela, pois nenhum artigo confirmou o uso deste procedimento. Os artigos estão ordenados por número de citações e ano de publicação.

Tabela 1: Artigos Selecionados

Ref.	Cit.	Ano	RNA	Reduzir taxa de erro	*N-gram*
[2]	49	2015	LSTM	Um único LSTM de duas camadas	Sim
[3]	33	2014	NNLM	Método para aproximar um NNLM com um LM de *back-off*	Sim
[4]	25	2013	NNLM	Utilizando um modelo variacional	Sim
[5]	19	2013	HMM/ANN	Combinando as pontuações de atributo com probabilidade de HMM	Sim
[6]	18	2013	NNLM	Processo ROVER e modelo acústico de adaptação cruzada	Sim
[7]	16	2013	RNNLM	Utilizando a RNNLM-Brown	Sim
[8]	15	2015	MTL-DNN	Dois métodos na estrutura de aprendizagem multitarefa	Não
[9]	10	2016	BRNN-LM	Rede neural recorrente bayesiana para LM	Sim
[10]	6	2017	DNN	Arquitetura DNN e uma técnica de otimização	Não
[11]	1	2017	DNN	Utilizando a abordagem DDA	Não
[12]	0	2017	DBRNN	Aplicando RNNs bidirecionais profundas	Sim
[13]	0	2017	DNN	Utilizando a técnica *phone-to-articulatory*	Não

3 ANÁLISE DOS RESULTADOS

Pode-se perceber que em 2017 obteve-se 33% dos artigos selecionados, com isso tem-se uma visão que este tema é atual. Com base nos artigos selecionados, verificou-se que as duas principais redes neurais utilizadas são: NNLM (*Neural Network Language Modeling*) e DNN (*Deep Neural Network*). Além disso, pode-se perceber que 66% utilizam o modelo estatístico *n-gram*, e o principal modelo utilizado é o *back-off*. Os demais não utilizam *n-gram* para melhorar as taxas de erros do reconhecimento de voz.

Ainda baseado nos artigos selecionados, todos apresentaram alguma solução para reduzir as taxas de erros do reconhecimento de voz. Pode-se perceber que, em muitos casos, está solução está associada ao uso, em conjunto, de redes neurais e modelos estatísticos *n-gram*. Além disso, nenhum artigo apresentou alguma maneira de disponibilizar o reconhecimento *offline* de voz para Android.

4 CONCLUSÕES E TRABALHOS FUTUROS

Considerando as limitações de reconhecimento *online* de voz, que essencialmente depende da Internet, conclui-se que poderá ser útil dispor deste recurso de modo *offline*, para dispositivos móveis. A principal contribuição deste trabalho é disponibilizar informações obtidas por meio de uma revisão sistemática da literatura, selecionando as melhores alternativas para a utilização de redes neurais, e também técnicas mais apropriadas para a redução da taxa de erros, que podem ser úteis para auxiliar a resolução de um problema em aberto.

A partir desta pesquisa, os trabalhos futuros estarão direcionados para a consolidação de uma abordagem racional para o projeto e implementação do reconhecimento *offline* de voz contínua em dispositivos móveis. Para atender aos requisitos de processamento, o próximo passo será realizar testes para identificar a melhor rede neural e o melhor modelo estatístico *n-gram* a ser aplicado. O objetivo é reduzir a taxa de erros e otimizar o processamento e uso de memória visando obter o menor consumo possível de bateria do *smartphone*.

REFERÊNCIAS

[1] V. F. S. Alencar. 2005. Atributos e Domínios de Interpolação Eficientes em Reconhecimento de Voz Distribuído. Master's thesis. Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brasil.

[2] M. Sundermeyer, H. Ney and R. Schlüter. 2015. From Feedforward to Recurrent LSTM Neural Networks for Language Modeling. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. IEEE.

[3] E. Arisoy, S. F. Chen, B. Ramabhadran and A. Sethy. 2014. Converting Neural Network Language Models into Back-off Language Models for Efficient Decoding in Automatic Speech Recognition. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. IEEE.

[4] A. Deoras, T. Mikolov, S. Kombrink and K. Church. 2013. Approximate inference: A sampling based modeling technique to capture complex dependencies in a language model. In *Speech Communication*. EURASIP/ISCA.

[5] S. M. Siniscalchi, T. Svendsen and C. Lee. 2013. A Bottom-Up Modular Search Approach to Large Vocabulary Continuous Speech Recognition. In *IEEE Transactions on Audio, Speech, and Language Processing*. IEEE.

[6] X. Liu, M. J. F. Gales and P.C. Woodland. 2013. Language model cross adaptation for LVCSR system combination. In *Computer Speech & Language*. ISCA.

[7] Y. Shi, W. Zhang, J. Liu and M. T. Johnson. 2013. RNN language model with word clustering and class-based output layer. In *Journal on Audio, Speech, and Music Processing*. EURASIP.

[8] D. Chen and B. K. Mak. 2015. Multitask Learning of Deep Neural Networks for Low-Resource Speech Recognition. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. IEEE.

[9] J. Chien and Y. Ku. 2016. Bayesian Recurrent Neural Network for Language Modeling. In *IEEE Transactions on Neural Networks and Learning Systems*. IEEE.

[10] A. L. Maas, P. Qi, Z. Xie, A. Y. Hannun, C. T. Lengerich, D. Jurafsky and A. Y. Ng. 2017. Building DNN acoustic models for large vocabulary speech recognition. In *Computer Speech & Language*. ISCA.

[11] S. Sun, B. Zhang, L. Xie and Y. Zhang. 2017. An unsupervised deep domain adaptation approach for robust speech recognition. In *Neurocomputing*. Elsevier.

[12] A. Ogawa and T. Horib. 2017. Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks. In *Speech Communication*. EURASIP/ISCA.

[13] B. Abraham and S. Umesh. 2017. An automated technique to generate phone-to-articulatory label mapping. In *Speech Communication*. EURASIP/ISCA.