Título: Progressos para a melhoria do reconhecimento automático da fala de crianças

Resumo alargado em Português

Nos últimos anos, o avanço considerável na tecnologia de reconhecimento automático da fala (*ASR*) proporcionou novas oportunidades em diversas áreas de aplicação, nomeadamente em aplicações direcionadas a crianças. A tecnologia *ASR* promete inovar as interações entre humanos e máquinas através do uso da fala, tutores de leitura automática e assistentes de patologia da fala, entre outras aplicações. No entanto, os desafios únicos colocados pela fala das crianças dificultam o desempenho dos atuais sistemas *ASR*, que foram concebidos para adultos. Por conseguinte, há uma necessidade urgente de soluções feitas à medida, adaptadas às características da fala das crianças. A investigação apresentada nesta tese dá passos importantes nesta direção, contribuindo com conhecimentos e avanços metodológicos importantes para o desenvolvimento de sistemas de *ASR* optimizados para essa população, assim abrindo novas possibilidades para a utilização da tecnologia de *ASR*.

Na fase inicial desta tese, fornecemos uma visão geral do contexto e dos desafios consideráveis associados ao reconhecimento da fala das crianças, estabelecendo deste modo os fundamentos do nosso problema de investigação. Sinteticamente, o desafio primário caracteriza-se pela significativa variabilidade presente nas componentes acústicas e linguísticas da fala das crianças, agravada pela variabilidade intra e inter-falantes. A resposta a este desafio multifacetado inerente ao reconhecimento do discurso infantil exige o acesso a um volume substancial de dados, o que constitui um segundo grande desafio. Enquanto os conjuntos de dados para a fala dos adultos são cada vez mais abundantes, estes dados em crianças continuam a ser escassos e geralmente limitados em tamanho. Nesta parte preliminar da tese, analisamos igualmente as diferentes metodologias empregues na literatura para abordar os diferentes desafios associados à *ASR* infantil. Através de uma análise aprofundada da literatura, identificamos as abordagens mais promissoras, que servem de base para o desenvolvimento desta tese. Adicionalmente, fornecemos uma compilação não exaustiva de corpora de fala infantil documentado na literatura, representando a coleção mais abrangente disponível até à data. Esta compilação constitui um recurso valioso para os investigadores, oferecendo uma visão dos conjuntos de dados existentes para a investigação e experimentação adicionais no contexto do reconhecimento do discurso infantil.

Na segunda parte da tese, centramo-nos na exploração de um sistema híbrido de *ASR* baseado em *Hidden Markov Models*, especificamente concebido para o reconhecimento da fala de crianças. Focamo-nos na língua inglesa, mas também numa língua com poucos recursos, especificamente o português europeu. Os nossos esforços centram-se na avaliação exaustiva da eficácia de várias abordagens de transferência de conhecimentos no domínio do ASR infantil. Das várias abordagens avaliadas, a aprendizagem por transferência surge como a técnica mais promissora para sistemas dedicados exclusivamente ao reconhecimento da fala infantil. A aprendizagem multitarefa, por outro lado, revela-se eficaz em cenários em que o sistema tem de reconhecer simultaneamente a fala de crianças e de adultos. Adicionalmente, introduzimos uma nova abordagem, designada "aprendizagem por transferência multilingue", que combina elementos das metodologias de aprendizagem multitarefa e de transferência. Os nossos resultados demonstram uma eficácia superior da

inicialização feita com o treino de um sistema *ASR* infantil multilingue, para a subsequente aprendizagem por transferência num conjunto de dados infantis, particularmente em contextos de poucos recursos. De facto, esta abordagem revela-se fundamental para atenuar os desafios associados à disponibilidade limitada de dados, abrindo caminho para sistemas de reconhecimento da fala infantil mais robustos e precisos em diversos contextos linguísticos.

Na fase subsequente desta tese, focamo-nos na exploração do paradigma *end-to-end*, com o objetivo de fazer avançar as atuais abordagens de *ASR* infantil. Partindo da abordagem convencional de transferência de aprendizagem sobre todo o modelo, propomos uma estratégia de avaliação com mais nuances. De facto, a nossa investigação revela o papel fundamental do codificador no processo de afinação do *ASR* infantil *end-to-end*. Esta descoberta vai de encontro à ideia de que, no contexto da fala das crianças, a variabilidade acústica supera significativamente os factores linguísticos, contribuindo para a degradação da precisão do reconhecimento. Para além disto, os nossos resultados evidenciam a eficácia da segmentação de camadas superiores, situadas mais perto da saída do codificador. Estes conhecimentos oferecem recomendações valiosas para otimizar o desenvolvimento de modelos ASR para crianças através da aprendizagem por transferência. Nesta secção da tese, também introduzimos o novo conceito de "*partial fine-tuning*" para arquiteturas baseadas em transformadores. As nossas descobertas indicam que o *fine-tuning* de componentes específicos da rede supera a abordagem tradicional de *fine-tuning* de todo o modelo. Nomeadamente, observamos que ajustar especificamente a componente *Feed-Forward Network* produz os melhores resultados. Esta abordagem inovadora é promissora no que toca a melhorar o desempenho dos sistemas *ASR* infantis, uma vez que concentra os esforços de afinação apenas nos components da rede com maior impacto no resultado final.

Em seguida, motivados pela necessidade de uma transferência de conhecimentos eficiente em termos de parâmetros, em particular em cenários com dados de treino limitados, investigamos a utilização de módulos *Adapters*. Estes módulos, compostos por duas camadas lineares integradas num modelo congelado pré-treinado, oferecem um mecanismo para a transferência de conhecimentos, mantendo os pesos e os conhecimentos encapsulados no modelo pré-treinado. A nossa investigação abrange igualmente a avaliação de várias configurações nas arquitecturas *Transformer* e *Conformer*. Entre as inúmeras configurações avaliadas, a configuração paralela, juntamente com a sua extensão *Conformer* conhecida como “T*wo Parallel Adapters*”, emerge como a melhor para transferir conhecimento para a fala das crianças. De forma notável, estsas configurações superaram o desempenho do ajuste fino de todo o modelo, alcançando resultados superiores usando apenas 10% dos parâmetros envolvidos no *transfer learning* tradicional. Esta avaliação enfatiza a promessa dos *Adapters* no contexto do ASR de crianças, e o seu potencial para uma adaptação mais precisa. Para melhorar ainda mais a adaptabilidade, introduzimos um procedimento não supervisionado em que os enunciados são agrupados utilizando o método *k-means* aplicado às respectivas representações de orador. Este método é justificado pela ideia de que a fala com características acústicas comparáveis, tal como detectada pelo extrator de incorporação do locutor, seria melhor se fosse utilizado um adaptador treinado em padrões de fala semelhantes, por oposição a um adaptador infantil geral.

De forma a expandir a eficácia dos adaptadores em colmatar a lacuna entre os domínios de origem (*source*) e de alvo (*target*) para a fala infantil, utilizamos os *Adapters* para melhorar o aumento de dados com dados imperfeitos para a ASR infantil. Especificamente, introduzimos o método inovador "*Double Way Adapter Tuning*" que, como estratégia para aumentar os dados disponíveis (*data augmentation*), utiliza a tecnologia de síntese de fala partir de texto (*text-to-speech (TTS)*). A nossa abordagem "*Double Way Adapter Tuning*" tem em conta as incompatibilidades acústicas entre o discurso sintético e o discurso real. Este método consiste num procedimento de duas etapas: inicialmente, treinar os módulos do *Adapters* utilizando dados *TTS* imperfeitos, seguido de um fine-tune dos *Adapters* e dos pesos de todo o modelo utilizando uma combinação de dados sintéticos e reais. Nomeadamente, os dados são submetidos a uma abordagem distinta de via dupla, com o discurso sintético a passar pelos *Adapters* enquanto o discurso real os contorna. A implementação da abordagem "*Double Way Adapter Tuning*" produz melhorias significativas em relação aos resultados de base e às técnicas anteriores nas arquitecturas *Transformer* e *Conformer*, destacando a eficácia do nosso método. Para além disto, alargamos a filtragem da *speaker embeddings* de dados sintéticos imperfeitos, incorporando x-vectors em vez de vectores. Isto envolve a utilização da semelhança de cosseno entre a referência e os enunciados gerados para descartar enunciados que possam ter sido gerados incorretamente, melhorando assim a qualidade e a fiabilidade dos dados sintéticos utilizados.

Inspirando-nos nos diversos sucessos observados com os *Adapters*, que abrangem tanto a transferência de *Adapters* como a abordagem inovadora "*Double Way Adapter Tuning*", avaliamos diferentes metodologias alternativas presentes na literatura. Os nossos resultados sublinham a eficácia duradoura dos adaptadores tradicionais como o método mais eficaz e eficiente em termos de parâmetros para melhorar o desempenho do *ASR* das crianças. A existência de um compromisso entre a precisão e a eficiência dos parâmetros é notada. Embora alguns métodos apresentem uma eficiência de parâmetros elevada, muitas vezes resultam em resultados significativamente degradados. Por outro lado, outras abordagens, embora menos eficientes em termos de parâmetros, produzem resultados comparáveis ou mesmo superiores ao ajuste fino de todo o modelo. Para atenuar este compromisso, introduzimos uma nova abordagem que aproveita a redundância inerente aos componentes *feed-forward* dos modelos baseados em transformadores. A nossa metodologia "*Shared-Adapters*", em que um único *Adapter* é partilhado por todas as camadas em vez de ser atribuído um por camada, demonstra um desempenho notável, superando o ajuste fino de todo o modelo. Apesar de enfrentar uma degradação mínima dos resultados de *ASR* em comparação com os *Adapters* tradicionais, o nosso adaptador partilhado é treinado com um número substancialmente menor de parâmetros do que qualquer abordagem anterior dentro desta gama de desempenho. Assim, a metodologia "*Shared-Adapters*" surge como uma excelente candidata para alcançar uma transferência de eficiência de parâmetros superior no contexto da *ASR* infantil, abrindo caminho para metodologias de adaptação mais eficazes.

Palavras-chave: Reconhecimento automático da fala, fala de criança, transferência de conhecimento, texto para fala, eficiência de parâmetros