Tecnologias de fala e a variação de pronúncia do russo no contexto de VoiceInteraction

Anna Havras[1,2]*, Carlos Mendes*[2]*, Gueorgui* Hristovsky[1]*, Sérgio Paulo*[2]*, Helena Moniz*[1,3]

[1]Universidade de Lisboa, Faculdade de Letras, Lisboa, Portugal
[2]VoiceInteraction – Tecnologias de Processamento de Fala, Lisboa, Portugal
[3]INESC-ID, Lisboa, Portugal

Resumo

O presente artigo tem como objetivo descrever o trabalho realizado na *VoiceInteraction*, empresa especializada no desenvolvimento de soluções de processamento de fala, com especial destaque para a transcrição automática, que recorre a um Reconhecedor Automático de Fala (ASR) híbrido. O objetivo principal centrou-se no estudo das características fonéticas da língua russa, tendo em conta quatro tarefas principais: descrição do inventário fonético-fonológico; validação das transcrições de noticiários; validação de um léxico previamente criado; e integração de pausas preenchidas no ASR. O presente trabalho contribuiu para o projeto *Artificial Intelligence and Advanced Data Analysis for Authority Agencies* (AIDA), financiado pela Comissão Europeia no âmbito do programa Horizonte 2020, transcrevendo os dados em língua russa.

Palavras-chave: reconhecimento automático de fala, fonética, pausas preenchidas, língua russa, variedades linguísticas.

Abstract

This article aims to describe the work conducted at *VoiceInteraction*, a company specialized in speech processing solutions, with a particular focus on automatic transcription using a Hybrid Automatic Speech Recognizer (ASR). The primary objective revolved around studying the phonetic characteristics of the Russian language, encompassing four main tasks: describing the phonetic-phonological inventory, validating news transcriptions, validating a previously created lexicon, and integrating filled pauses into the ASR. This work contributed to the Artificial Intelligence and Advanced Data Analysis for Authority Agencies (AIDA) project, funded by the European Commission under the Horizon 2020 program, by transcribing the data in the Russian language.

Keywords: automatic speech recognition, phonetics, filled pauses, Russian language, linguistic varieties.

1. Introdução

O presente trabalho consistiu na validação de um sistema de reconhecimento de fala automático, previamente criado, para a **língua russa**. Esta língua é falada em diferentes regiões dos continentes europeu e asiático. Assim, o objetivo foi o de identificar as variantes fonéticas e variedades linguísticas da língua russa e incorporá-las no reconhecedor, aumentando assim a qualidade do reconhecimento. Foi necessário realizar uma descrição detalhada do sistema fonético-fonológico da língua russa para a validação de um *phoneset* previamente criado e validação do modelo de léxico, que continha as palavras e as suas pronúncias correspondentes. As regiões analisadas foram: a Rússia Europeia, a Bielorrússia e o Cáucaso. Na Rússia Europeia, o russo é a língua oficial; na Bielorrússia, o russo é uma das línguas oficiais do país; e no Cáucaso, o

russo é usado como língua franca, visto que esta era falada na União Soviética e continua até hoje a ser utilizada pelos falantes.

Desta forma, para a validação de um **Reconhecedor Automático de Fala** (ASR) previamente criado para a língua russa, foi necessário recolher dados de fala e de texto de conteúdo noticioso das regiões referidas anteriormente. Os dados de fala foram transcritos automaticamente e editados pelos anotadores humanos, seguindo as orientações de anotação criadas pela empresa, a fim de corrigir os erros presentes nas transcrições automáticas.

Por outro lado, analisámos e validámos um conjunto de **pausas preenchidas**, previamente criado para a língua russa na empresa. Centrámos a nossa atenção na validação das pausas preenchidas, porque se revelaram um tipo de disfluência mais frequentemente utilizado pelos falantes nas transcrições. Observámos que os anotadores humanos apresentaram alguma dificuldade em identificá-las, sendo que o trabalho efetuado corresponde a uma primeira validação do conjunto das pausas preenchidas da língua. A falta de anotação das pausas preenchidas pelos transcritores humanos pode ter um impacto significativo no desempenho do ASR, quando aplicado a tarefas em que a fala é menos preparada e, por isso, estas disfluências são mais frequentes.

Por fim, analisámos o **léxico** da língua russa. O objetivo foi validar as pronúncias geradas automaticamente para as 10 mil palavras mais frequentes de um léxico composto por 400 mil entradas. Esta tarefa permitirá a extração de padrões linguísticos que serão utilizados no módulo *Grapheme-to-Phone* (G2P).

O presente trabalho contribuiu ainda para o projeto *Artificial Intelligence and Advanced Data Analysis for Authority Agencies* (AIDA), financiado pela Comissão Europeia no âmbito do programa Horizonte 2020. Após a validação do sistema de reconhecimento de língua russa, foi possível aplicar o presente trabalho ao AIDA, através da transcrição e da validação dos dados de conteúdo sensível de língua russa.

2. Fonética da língua russa

Para a validação do modelo de léxico e de pronúncia foi necessário recolher regras e regularidades da pronúncia da língua russa. A informação apresentada neste capítulo é uma seleção dos trabalhos de Litnevskaya (2006), Demidov et al. (2013), Kasatkin (2014), Popov (2014), Yanushevskaya e Bunčić (2015), Osipova (2018) e Sokolova (2021).

De acordo com Jurafsky e Martin (2021), a fonética é o estudo dos sons da fala utilizados nas línguas do mundo, como são produzidos no trato vocal humano, como são realizados acusticamente e como podem ser digitalizados e processados. A Tabela 1 apresenta o sistema fonético da língua russa. Na primeira coluna são apresentadas as variantes fonéticas de acordo com o Alfabeto Fonético Internacional (IPA); na segunda coluna, as mesmas variantes são apresentadas de acordo com o X-SAMPA; na terceira coluna, são apresentados os possíveis grafemas correspondentes a cada uma das variantes fonéticas; e finalmente, na última coluna, exemplos com transcrições fonéticas. As vogais são apresentadas no lado esquerdo da tabela, e as consoantes no lado direito.

Tabela 1. Sistema fonético da língua russa

IPA	X-SAMPA	Grafemas	Exemplo	IPA	X-SAMPA	Grafemas	Exemplo
I	i	И	*Ми**л**а* [mʲ**i**ɬə]	p	p	п	***п**ока* [**p**ɐká]
ɪ	I	и, е, э, а, я	*м**е**док* [mʲ**ɪ**dók]	pʲ	p'	п	***п**ятак* [**pʲ**ɪták]
E	e	Е	*м**е**ра* [mʲ**é**ra]	b	b	б	***б**уря* [**b**úrʲə]
ɛ	E	Э	*м**э**р* [m**ɛ́**r]	bʲ	b'	б	***б**юро* [**bʲ**uró]
A	a	а, я	*м**а**л* [m**á**ɬ]	t	t	т	*э**т**аж* [ɪ**t**áʂ]
ɐ	6	а, о	*ж**а**ра* [ʐ**ɐ**rá]	tʲ	t'	т	*жи**т**ь* [ʐɨ**tʲ**]
ə	@	а, о, я	*бур**я*** [búrʲ**ə**]	d	d	д	*ря**д**ы* [rʲɪ**d**ɨ]
Æ	{	я, а	*ч**а**й* [t͡ɕʲ**æ**j]	dʲ	d'	д	***д**ень* [**dʲ**énʲ]
ɨ	1	ы, и, а, е	*м**ы**ло* [m**ɨ**ɬə]	k	k	к	*уро**к*** [ʊró**k**]
ɔ	O	О	*как**о*** [kɐká**ɔ**]	kʲ	k'	к	*ле**к**ий* [lʲóxʲ**kʲ**ɪj]
O	o	О	*м**о**л* [m**ó**ɬ]	g	g	г	*и**г**ра* [ɪ**g**rá]
ɵ	8	ё, о	*вахт**ё**ры* [vɐxtʲ**ó**rɨ]	gʲ	g'	г	*серь**г**и* [sʲɪrʲ**gʲ**ɪ]
U	u	у, ю	*б**у**ря* [b**ú**rʲə]	x	x	х	***х**леб* [**x**lʲép]
ʊ	U	у, ю	*авг**у**ст* [ávg**ʊ**st]	xʲ	x'	х	*арх**и**в* [ɐr**xʲ**íf]
ʉ	}	у, ю	*зл**ю**сь* [zlʲ**ʉ́**sʲ]	f	f	ф	*алло**ф**он* [ɐlɐ**f**ón]
				fʲ	f'	ф	***ф**ильм* [**f**ʲílʲm]
				v	v	в	***в**озди* [**v**ózʲdʲɪ]
				vʲ	v'	в	***в**юга* [**vʲ**úgə]
				s	s	с	*ча**с**ы* [t͡ɕʲɪ**s**ɨ]
				ʐ	z`	ж	***ж**ара* [**ʐ**ɐrá]
				t͡s	ts	ц	***ц**ель* [**t͡s**élʲ]
				t͡ɕ	ts\'	ч	***ч**ай* [**t͡ɕ**áj]
				ɕː	s\:'	щ	***щ**ётка* [**ɕː**ʲɵ́tkə]
				j	j	й	*пусто**й*** [pʊstó**j**]
				ɬ	5	л	*мо**л*** [mó**ɬ**]
				lʲ	l'	л	*ате**л**ье* [ɐtʲ**lʲ**jé][a]
				m	m	м	*съё**м**ка* [sʲjó**m**kə]
				mʲ	m'	м	***м**ера* [**mʲ**éra]
				n	n	н	*пья**н*** [pʲjæ**n**]
				nʲ	n'	н	*де**н**ь* [dʲé**nʲ**]
				r	r	р	*ме**р**а* [mʲé**r**a]
				rʲ	r'	р	*бу**р**я* [bú**rʲ**ə]

[a] Esta palavra também pode ser pronunciada como [ɐtɛlʲjé], com a vogal [ɛ] ao invés de [ɪ] na segunda sílaba.

2.1. **Pronúncia das vogais**

Nesta secção, apresentamos 15 das variantes fonéticas mais frequentes das vogais em russo. Cada uma é apresentada de acordo com os seus possíveis grafemas e exemplos.

As vogais que podem ocorrer em **posição tónica** na língua russa são [a], [æ], [i], [ɨ], [e], [ɛ], [o], [ɔ], [ɵ], [u] e [ʉ]. A Tabela 2 mostra as variantes fonéticas que podem ocorrer em posição tónica e os seus possíveis grafemas correspondentes, seguidos de exemplos, transcrição fonética e o significado equivalente em português.

Tabela 2. Vogais tónicas e os seus grafemas correspondentes (adaptado de Litnevskaya, 2006)

Vogais acentuadas	Grafemas	Palavra	Pronúncia	Significado
[a]	а - [a]	*м**а**л*	[m**á**ɫ]	‘pequeno’
	я - [ʲa]	*м**я**л*	[mʲ**á**ɫ]	‘amassado’
[æ]	а - [a]	*ч**а**й*	[t͡ɕʲ**ǽ**j]	‘chá’
	я - [ʲa]	*п**я**льцы*	[pʲ**ǽ**lʲt͡sɨ]	‘aro
[ɨ]	и - [i]	*М**и**ла*	[mʲ**ɨ́**ɫə]	nome próprio ‘Mila’
[ɨ]	и - [i]	*ж**и**ть*	[ʐ**ɨ́**tʲ]	‘viver’
	ы - [ɨ]	*м**ы**ло*	[m**ɨ́**ɫə]	‘sabão’
[e]	е - [ʲe]	*м**е**ра*	[mʲ**é**ra]	‘medida’
[ɛ]	э - [ɛ]	*м**э**р*	[m**ɛ́**r]	‘Presidente da Câmara’
[o]	о - [o]	*м**о**л*	[m**ó**ɫ]	‘cais’
[ɔ]	о - [o]	***о**кна*	[**ɔ́**knə]	‘janelas’
[ɵ]	о - [o]	*щипач**о**м*	[ɛːʲɪpʁt͡ɕʲ**ɵ́**m]	‘com a pinça’
	ё - [ʲo]	*вахт**ё**ры*	[vɐxtʲ**ɵ́**rɨ]	‘porteiros’
[u]	у - [u]	*б**у**ря*	[b**ú**rʲə]	‘tempestade’
	ю - [ʲu]	*со**ю**з*	[sɐj**ú**z]	‘união’
[ʉ]	у - [u]	*щ**у**рить*	[ɛːʲ**ʉ́**rʲɪtʲ]	‘piscar os olhos’
	ю - [ʲu]	*зл**ю**сь*	[zlʲ**ʉ́**sʲ]	‘Eu zango-me’

Em posição átona podem ocorrer as seguintes variantes fonéticas - [ɐ, ə, ɪ, ɨ, ɔ, ʊ, ʉ]. Os exemplos a seguir representam as regularidades em posições átonas descritas para esta língua.

Após as consoantes fortes (exceto [ʂ], [ʐ] e [t͡ʂ]) podem ser encontrados grafemas como *y* - [u] realizado como [ʊ] (1a); *a* - [a] e *o* - [o] realizados como [ɐ] / [ə] (1b-1e); *ы* - [ɨ] e *e* - [ʲe] realizados como [ɪ] (1f, 1g):

(1a) *y*	[ʊ]	*п**у**стой*	[p**ʊ**stój]	‘vazio’
(1b) *a*	[ɐ]	*с**а**ма*	[s**ɐ**má]	‘sozinha’
(1c) *a*	[ə]	*камер**а***	[kámʲɪr**ə**]	‘câmara’
(1d) *o*	[ɐ]	*п**о**стель*	[p**ɐ**sʲtʲélʲ]	‘cama’
(1e) *o*	[ə]	*г**о**лова*	[g**ə**ɫɐvá]	‘cabeça’
(1f) *ы*	[ɪ]	*м**ы**слитель*	[m**ɪ**slʲítʲɪlʲ]	‘pensador’
(1g) *e*	[ɪ]	*т**е**стировать*	[t**ɪ**stʲírəvətʲ]	‘testar’

Após as variantes [ʂ], [ʐ] e [t͡ʂ], podem ser encontrados grafemas como *y* - [u] realizado como [ʊ] (2a); *a* - [a] e *o* - [o] realizados como [ɐ], [ɨ] ou [ə] (2b-2d); *ы* - [ɨ], *e* - [ʲe] e *u* - [i] realizados como [ɨ] (2e-2h):

(2a) *у*	[ʊ]	***щу**меть*	[ɕʊmʲétʲ]	‘fazer barulho’
(2b) *а*	[ɐ]	*ж**а**ра*	[zɐ̱rá]	‘calor’
(2c) *а*	[ɨ]	*лош**а**дей*	[łəɕɨdʲéj]	‘dos cavalos’
(2d) *а*	[ə]	*мужж**а***	[múzə̱]	‘do marido’
(2e) *о*	[ɨ]	*ш**о**колад*	[ɕɨkɐłát]	‘chocolate’
(2f) *ы*	[ɨ]	*ножниц**ы***	[nózn̪ʲɪt͡sɨ̱]	‘tesoura’
(2g) *и*	[ɨ]	*тиш**и**ною*	[tʲɪɕɨnójʊ]	‘com o silêncio’
(2h) *е*	[ɨ]	*ж**е**лезо*	[zɨ̱lʲézə]	‘ferro’

Após as consoantes suaves, podem ser encontrados grafemas como *у* - [u] e *ю* - [ʲu] realizados como [ʊ]/ [ʉ] (3a-3d); *а* - [a], *и* - [i], *е* - [ʲe] e *я* - [ʲa] realizados como [ɪ] (3e-3h):

(3a) *у*	[ʊ]	*ч**у**десный*	[t͡ɕʲʊ̱dʲésnɨj]	‘adorável’
(3b) *ю*	[ʊ]	*л**ю**бить*	[lʲʊ̱bʲɪtʲ]	‘amar’
(3c) *у*	[ʉ]	*ощ**у**щение*	[ɐɕːʲʉ̱ɕːʲenʲɪje]	‘sentimento’
(3d) *ю*	[ʉ]	*следу**ю**щий*	[sʲlʲédʊjʉ̱ɕːʲɪj]	‘seguinte’
(3e) *а*	[ɪ]	*ч**а**сы*	[t͡ɕʲɪ̱sɨ́]	‘horas’
(3f) *и*	[ɪ]	*м**и**ры*	[mʲɪ̱rɨ́]	‘mundos’
(3g) *е*	[ɪ]	*м**е**нять*	[mʲɪ̱nʲátʲ]	‘mudar’
(3h) *я*	[ɪ]	*п**я**так*	[pʲɪ̱ták]	‘níquel’

Por fim, no início da palavra, podem ser encontrados grafemas como *у* - [u] realizado como [ʊ] (4a); *а* - [a] e *о* - [o] realizados como [ɐ] (4b, 4c); *и* - [i] realizado como [ɪ] (4d); e э - [ɛ] realizado como [ɛ], [ɪ] e [ɨ] (4e):

(4a) *у*	[ʊ]	***у**рок*	[ʊ̱rók] ‘aula’
(4b) *а*	[ɐ]	***а**ллофон*	[ɐ̱łɐfón] ‘alofone’
(4c) *о*	[ɐ]	***о**кно*	[ɐ̱knó] ‘janela’
(4d) *и*	[ɪ]	***и**гра*	[ɪ̱grá] ‘jogo’
(4e) э	[ɪ]	***э**таж*	[ɛ̱táʐ]/ [ɪ̱táʐ]/ [ɨ̱táʐ] ‘andar’

2.2. **Pronúncia das consoantes**

Na língua russa, as consoantes são organizadas em pares de fortes-suaves, com exceção de [ɕ, ʐ, t͡s] - que são apenas fortes, e [t͡ɕʲ, ɕːʲ, j] - que são apenas suaves. Podemos ver todas as consoantes fortes/suaves na Tabela 3.

Tabela 3. Divisão das consoantes em fortes-suaves e pares-ímpares (adaptado de Litnevskaya, 2006)

Consoantes	**Par**	**ímpar**
fortes	[b, v, g, d, z, k, ł, m, n, p, r, s, t, f, x]	[ɕ, ʐ, t͡s]
suaves	[bʲ, vʲ, gʲ, dʲ, zʲ, kʲ, lʲ, mʲ, nʲ, pʲ, rʲ, sʲ, tʲ, fʲ, xʲ]	[t͡ɕʲ, ɕːʲ, j]

As consoantes suaves distinguem-se das consoantes fortes devido a uma articulação adicional - a **palatalização** que se sobrepõe à articulação básica de uma consoante.

Da mesma forma, existe uma divisão das consoantes em vozeadas - pronunciadas com a vibração das cordas vocais e surdas - pronunciadas sem qualquer vibração das cordas vocais. A Tabela 4 apresenta as diferentes consoantes divididas em vozeadas, surdas e pares/ímpares.

Tabela 4. Divisão das consoantes em vozeadas-surdas e pares-ímpares (adaptado de Litnevskaya, 2006)

Consoantes	**par**	**ímpar**
vozeadas	[b, bʲ, v, vʲ, g, gʲ, d, dʲ, z, zʲ, ʐ]	[j, ɬ, lʲ, m, mʲ, n, nʲ, r, rʲ]
surdas	[p, pʲ, f, fʲ, k, kʲ, t, tʲ, s, sʲ, ʂ, x, xʲ]	[t͡s, t͡ɕ, ɕː]

A distinção das **consoantes fortes *versus* suaves** pode ser feita através da grafia. No russo, o [j] em superescrito nas vogais [ʲa], [ʲu], [ʲo] e [ʲe] indica uma propriedade de uma consoante precedente que é suave. Vogais como [i] e [ɪ] também indicam que a consoante precedente é suave:

(5a)	***мал***	[**m**áɬ]	‘pequeno’	*versus*	***мял***	[**m**ʲáɬ]	‘ele amassou’
(5b)	***мол***	[**m**óɬ]	‘cais’		***мёл***	[**m**ʲóɬ]	‘giz’
(5c)	***пэр***	[**p**ér]	‘par’		***перо***	[**p**ʲeró]	‘caneta’
(5d)	***буря***	[**b**úrʲa]	‘tempestade’		***бюро***	[**b**ʲʉró]	‘escritório’
(5e)	***мыло***	[**m**ɨ́ɬə]	‘sabão’		***мило***	[**m**ʲíɬə]	‘agradável’

O sinal gráfico <ь>[1] indica, também, que a consoante precedente par é suave e ocorre em:

— no fim de uma palavra - *конь* [kó**n**ʲ] ‘cavalo’;

— no meio de uma palavra, depois da consoante [l] e seguido por qualquer outra consoante - *полька* [pó**l**ʲkə] ‘dança polonesa’;

— depois de uma consoante vozeada e antes de uma consoante surda - *весьма* [vʲɪ**s**ʲmá] ‘muito’;

— e quando uma consoante suave é seguida por uma das consoantes [gʲ], [kʲ], [bʲ] e [mʲ] - *серьги* [sʲɪ**r**ʲgʲɪ́] ‘brincos’.

Para além disso, quando o sinal gráfico <ь> se encontra antes de [ʲa], [ʲu], [ʲo] e [ʲe], indica que a consoante precedente é suave, mas não suaviza a própria consoante (a consoante e a vogal ocorrem em sílabas separadas, uma vez que o segmento palatalizado [j] se encontra na posição de ataque):

(6a)	***пьян***	[**p**ʲ**j**án] ‘bêbedo’
(6b)	***вьюга***	[**v**ʲ**j**ú**g**ə] ‘nevão’
(6c)	*бульон*	[bʊ**l**ʲ**j**ó**n**] ‘caldo’
(6d)	*ателье*	[ɐtʲɪ**l**ʲ**j**é] ‘Atelier’

Embora a ortografia não nos forneça informações sobre certas consoantes, o facto de elas ocorrerem em certas posições significa que sabemos como pronunciá-las. Isso acontece em contextos como:

[1] O sinal gráfico <ь> também tem a função de marcador de infinitivo quando ocorre em verbos infinitivos e não suaviza a consoante precedente - *учиться* [ʊt͡ɕɪ́**ts**ə] ‘estudar’.

— [n] tem a pronúncia de [nʲ] quando a consoante seguinte é um [t͡ɕʲ] ou [ɕːʲ]:

(7a) *бараба**н**чик* [bərʁbá**n**ʲt͡ɕʲɪk] ‘bateria’
(7b) *бараба**н**щик* [bərʁbá**n**ʲɕːʲɪk] ‘baterista’[2]

— [s] tem a pronúncia de [sʲ] quando a consoante seguinte é [nʲ] ou [tʲ]:

(8a) *ко**с**ть* [kó**s**ʲtʲ] ‘osso’
(8b) *пе**с**ня* [pʲé**s**ʲnʲə] ‘canção’

— [z] tem a pronúncia de [zʲ] quando a consoante seguinte é [nʲ] ou [dʲ]:

(9) *жи**з**нь* [ʐʲí**z**ʲnʲ] ‘vida’

— os fones alveolares dentais fortes, exceto [n] e [ɬ], que precedem os fones labiodentais suaves, podem sofrer um processo de palatalização[3]:

(10a) ***д**верь* [**d**ʲvʲérʲ] / [**d**vʲérʲ] ‘porta’
(10b) *ра**з**ве* [rá**z**ʲvʲe] / [rá**z**vʲe] ‘é’

Em oposição, as **consoantes fortes pares** podem ser identificadas através de:

— ausência do sinal gráfico <ь> que indica que a consoante precedente é forte (não palatalizada):

(11) ба**н**ка [bá**n**kə] ‘jarro’

— quando seguidas de vogais [a, o, ɛ, u, ɨ,][4]:

(12a) ***м**ал* [**m**áɬ] ‘pequeno’
(12b) ***м**ол* [**m**óɬ] ‘cais’
(12c) ***п**эр* [**p**ér] ‘par’
(12d) ***б**уря* [**b**úrʲa] ‘tempestade’
(12e) ***м**ыло* [**m**íɬə] ‘sabão’

— também, o sinal gráfico <ъ> (distinto de <ь>) antes de [ʲa], [ʲu], [ʲo] e [ʲe] indica que a consoante precedente é forte:

(13a) *об**ъя**ть* [ʋ**bjá**tʲ] ‘abraçar’
(13b) *ко**нъю**нкт* [kʋ**njú**nkt] ‘conjunto’

[2] Isto também pode ser explicado através do processo de realização/não-realização dos yers nas línguas eslavas. Os yers alternam com o zero nas sílabas iniciais das palavras, nas sílabas finais das palavras, nas posições prefixais e no meio da palavra. Quando o yer não se reflete na ortografia, pode ainda assim influenciar a pronúncia, afetando a consoante anterior. Para mais pormenores sobre este assunto, consulte Yearley (1995).
[3] De acordo com o *Dicionário Ortográfico* de Borunova, Vorontsova, Eskova e Avanesov (1989), os fones alveolares dentais fortes, exceto [n] e [ɬ], que precedem os sons labiodentais suaves, passam obrigatoriamente por um processo de palatalização. Atualmente, esta realização é rara, e o som forte é preferido.
[4] Em alguns estrangeirismos, uma consoante forte pode ser seguida de [ʲe] - *б**е**кон* [bʲ**e**kón] ‘bacon’.

(13c)	***адъюнкт***	[ɐ**dju**nkt]	‘adjunto’
(13d)	***съест***	[**sjé**st]	‘vai comer’
(13e)	***съёмка***	[**sjó**mkə]	‘sessão’

2.2.1. **Assimilação das consoantes**

Em russo, na grafia existem sequências de duas obstruintes consecutivas. Para simplificar a pronúncia, muitas vezes o primeiro segmento assimila a articulação ou as propriedades acústicas da consoante seguinte, resultando em duas consoantes da mesma qualidade e, consequentemente, na assimilação completa, apagando uma das consoantes.

Podemos observar na Tabela 5, os exemplos de assimilação de consoantes, extraídos de Litnevskaya (2006).

Tabela 5. Assimilação consonântica (adaptado de Litnevskaya, 2006)

Grafemas	**Assimilação das consoantes**				**Significado**
с + ш	[s] + [ʂ]	[ʂː]	***сш****ить*	[**sː**ɨtʲ]	‘costurar’
з + ж	[z] + [ʐ]	[ʐː]	***изж****ить*	[ɪ**ʐː**ítʲ]	‘livrar-se de’
т + ц	[t] + [ts]	[tsː]	***отц****епить*	[ɐ**tsː**ɪpʲítʲ]	‘desengatar’
т + ч	[t] + [tɕ]	[tɕː]	***отч****ет*	[ɐ**tɕː**ót]	‘relatório’
с + щ	[s] + [ɕː]	[ɕː]	*расщ****е****пить*	[rə**ɕː**ɪpʲítʲ]	‘dividir’
с + ч	[s] + [tɕ]	[ʂːʲtɕ]	***с ч****ем-то*	[**sːʲtɕ**émtə]/ [**sː**ʲémtə]	‘com alguma coisa’
т + с	[t] + [s]	[ts]	*мы****тьс****я*	[mɨ**ts**ə]	‘lavar-se’
т + щ	[t] + [ɕː]	[tɕɕː]	***отщ****епить*	[ɐ**tɕɕː**ɪpítʲ]	‘descolar’

Na primeira coluna da tabela, são apresentados os grafemas que serão assimilados, na segunda coluna, a respetiva realização desses grafemas e, na terceira coluna, a realização das consoantes após a assimilação.

Existem também contextos de **assimilação simples**, na qual uma consoante assimila apenas o vozeamento, como em *бу****д****ка* <bu**d**ka> que se transforma em [bú**t**kə] ‘cabine’. A consoante [d] assimila o traço de [-vozeado] da consoante seguinte [k], resultando em [t]. As consoantes surdas pares antes das consoantes vozeadas (exceto [v, vʲ, jʲ, ł, lʲ, m, mʲ, n, nʲ, r, rʲ]) transformam-se em vozeadas, pois assimilam o traço da consoante seguinte, como em *моло****т****ьба* <małat**b**á> → [məłɐ**d**ʲbá] ‘espancamento’ e ***с****дела****т****ь* <**s**délat> → [**z**ʲdʲélətʲ] ‘fazer’.

Além disso, existe um fenómeno de **neutralização da obstruinte** em posição final de palavra:

— a consoante /v/ é realizada como [f], como em *архи****в*** <arxi**v**> → [ɐrxʲí**f**] ‘arquivo’

— a consoante /b/ é realizada como [p], como em *хле****б*** <xlé**b**> → [xlʲé**p**] ‘pão’

— a consoante /z/ é realizada como [s], como em *указ* <ukáz> → [ʊká**s**] ‘decreto’

— a consoante /ʐ/ é realizada como [ʂ], como em *экипа****ж*** <ekipá**ʐ**] → [ɛkʲɪpá**ʂ**] / [ɪkʲɪpá**ʂ**] / [ɨkʲɪpá**ʂ**] ‘tripulação’.

Ao contrário da assimilação, existe na língua russa o fenómeno de **dissimilação consonântica**. Podemos observar este fenómeno fonológico na presença de [g] mais [k], em que [g] dissimila-se de [k] e transforma-se em [x], como em *легко* <le**g**ko> → [lʲí**x**kə] ‘facilmente’ ou *мягкий* <ma**g**kij> → [mʲæ**x**kʲɪj] ‘suave’.

2.2.2. **Apagamento das consoantes**

Esta secção apresentará um conjunto de estruturas em que a combinação de certas consoantes favorece o apagamento de uma delas quando a palavra é realizada. De acordo com Litnevskaya (2006), há casos em que

uma palavra na grafia apresenta três consoantes consecutivas, mas, quando pronunciada, uma delas é apagada (normalmente uma consoante que se encontra no meio). Este processo resulta em sílabas separadas, com uma consoante em posição de coda e a outra em posição de ataque.

Tabela 6. Apagamento das consoantes (adaptado de Litnevskaya, 2006)

Grafemas	**Apagamento das consoantes**				**Significado**
стл	[stɬ]	[sɬ]	*сча**стл**ивый*	[ɕːʲ**sl**ívɨj]	‘feliz’
стн	[stn]	[sn]	*ме**стн**ый*	[mʲé**sn**ɨj]	‘local’
здн	[zdn]	[zn]	*по**здн**ий*	[pózʲ**n**ʲij]	‘atrasado’
здц	[zd͡ts]	[s͡ts]	*под у**здц**ы*	[pʁdʊ**s͡ts**ɨ]	‘sob as rédeas’
ндш	[ndʂ]	[nʂ]	*ла**ндш**афт*	[ɬʁ**n**sáft]	‘paisagem’
нтг	[ntg]	[ng]	*ре**нтг**ен*	[rʲɪ**ng**ʲén]	‘Raio-X’
ндц	[nd͡ts]	[n͡ts]	*голла**ндц**ы*	[gʁɬá**n͡ts**ɨ]	‘holandeses’
рдц	[rd͡ts]	[rt͡s]	*се**рдц**е*	[sʲé**rt͡s**ɨ]	‘coração’
рдч	[rd͡tɕʲ]	[rt͡ɕʲ]	*се**рдч**ишко*	[sʲɪ**rt͡ɕ**íʂkə]	‘coração pequeno
лнц	[ɬn͡ts]	[n͡ts]	*со**лнц**е*	[só**n͡ts**ə]	‘sol’
V1[ɪ]	[V1jɪ]	[V1ɪ]	*мое**го***	[mə**ɪ**vó]	‘do meu’

Na primeira coluna da Tabela 6, podemos observar os grafemas e, na segunda coluna, a sua pronúncia correspondente, na terceira coluna é apresentada a pronúncia após o apagamento de uma das consoantes.

3. **Disfluências**

Como Moniz (2013) mencionou, as disfluências são uma parte crucial da fala humana espontânea (Levelt, 1983; Allwood et al., 1990; Swerts, 1998; Clark e Fox Tree, 2002). Estas são usadas na estruturação do discurso para planear o discurso subsequente e podem introduzir informação nova (Clark e Fox Tree, 2002; Arnold et al., 2003).

Os falantes filtram automaticamente as disfluências, o que parece ser um desafio para os sistemas ASR. Segundo Shriberg (2001, p. 153):

> “There is good reason to study disfluencies, in both theoretical and applied fields. They are frequent - affecting up to ten percent of words and over one third of utterances in natural conversation. For the study of human language, disfluencies provide a window onto underlying processes affecting human speech and language production. On the applied side, disfluencies present a challenge for automatic speech processing, especially since speech recognition models are often trained on read or highly constrained speech (Butzberger et al., 1992).”

Como discutido em Schettino (2022), para melhorar o reconhecimento automático da fala humana espontânea, alguns estudos propuseram inicialmente filtrar as disfluências, detetando-as e cortando-as das transcrições (Bear et al., 1992; Gabrea & O'Shaughnessy, 2000; Heeman & Allen, 1994; Nakatani & Hirschberg, 1994; Oviatt, 1995). Desta forma, os resultados seriam transcrições "limpas" e "perfeitas", mas longe da produção. Consequentemente, a investigação sobre disfluências começou a ganhar interesse, uma vez que se trata de um mecanismo natural da fala espontânea e é uma parte crucial do processamento natural. Na ASR, os investigadores começaram a desenvolver modelos incluindo disfluências, que melhoraram o reconhecimento e o processamento do discurso espontâneo (Heeman & Allen, 2001; Hough & Purver, 2012; Liu et al., 2006). Assim, as disfluências deixaram de ser consideradas um obstáculo, mas um mecanismo linguístico que precisava de ser estudado e incluído na ASR em vez de ser simplesmente eliminado. A sua

inclusão também melhorou a interação homem-máquina, uma vez que é mais natural e assemelha-se à fala humana (Adell et al., 2012; Betz, 2020).

Acreditamos que a investigação sobre as pausas preenchidas na língua russa pode melhorar o sistema de reconhecimento da VoiceInteraction, reduzindo o tempo de processamento e aumentando a qualidade, uma vez que as disfluências, em geral, e as pausas preenchidas, em particular, afetam não só a palavra a ser reconhecida, mas também as sequências adjacentes.

3.1. **Pausas preenchidas**

O presente trabalho focou-se no estudo das pausas preenchidas, visto que estas são o tipo de disfluência mais frequentemente utilizado pelos falantes nos dados de fala analisados, e os anotadores humanos apresentaram algumas dificuldades na identificação das mesmas. Tal como estudado por Barczewska e Igras (2012), a frequência das pausas preenchidas depende em grande medida dos falantes e das tarefas a realizar. No entanto, um falante pode atingir até 10 pausas preenchidas por minuto. Uma vez que os anotadores humanos não transcrevem as pausas preenchidas, a não identificação das pausas causa problemas no ASR.

As pausas preenchidas são usadas pelos falantes para sinalizar reparações e planear as unidades seguintes, e são marcas de trabalho de formulação. São utilizadas em situações de pesquisa lexical, por exemplo, quando o falante ainda está a pensar na palavra correta a produzir, sendo frequente produzir uma pausa preenchida em vez de uma pausa silenciosa. As pausas preenchidas são utilizadas na interação comunicativa para manter ou retomar a palavra, e podem variar em duração, apresentando produção mais longa ou mais curta na formulação do discurso (Moniz, 2013).

Kharlamova (2008) estudou as pausas preenchidas da língua russa através da análise de questionários preenchidos por falantes nativos de russo, análise acústica do discurso na rádio, na televisão e em conferências, e observações de fala em comunicação direta. O estudo revelou as seguintes pausas preenchidas: [e], [ɨ], [m], [n], e [n']. Teng (2015) realizou um estudo de fala espontânea recolhida de sujeitos russos, chineses e americanos, para demonstrar pausas preenchidas universais e específicas de cada língua. Através da medição dos formantes e da observação do espectrograma, as pausas preenchidas utilizadas predominantemente pelos falantes de russo são os sons [a], [am] e [m]. O trabalho de Bogdanova-Belgarian e Baeva (2018) teve como objetivo listar vários elementos não verbais do The Speech Corpus of the Russian Language, entre os quais as pausas preenchidas. Segundo as autoras, além das pausas preenchidas descritas acima, sons [ə] e [əm] também são frequentemente utilizados pelos falantes de russo como pausas preenchidas.

Desta forma, as pausas preenchidas do russo, segundo a literatura, são as apresentadas na Tabela 7. Estas são precedidas de símbolo de percentagem " % ", visto que é desta forma que são sinalizadas pelos anotadores humanos numa transcrição na VoiceInteraction.

Tabela 7. Pausas preenchidas da língua russa de acordo com Kharlamova (2008), Teng (2015), Bogdanova-Belgarian e Baeva (2018) e Teng e Androsova (2022)

Russo	**IPA**	**X-SAMPA**
%а	[a]	[a]
%ам	[am]	[am]
%@	[ə]	[@]
%@м	[əm]	[@m]
%э	[e]	[e]
%ы	[ɨ]	[1]
%м	[m]	[m]
%н(ь)	[n] / [nʲ]	[n] / [n']

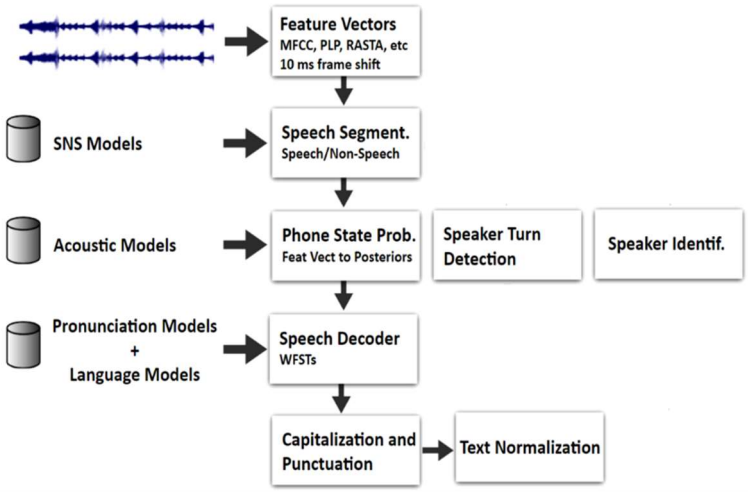
4. Sistema de Reconhecimento Automático de Fala

As tarefas desenvolvidas ao longo do presente trabalho foram aplicadas ao sistema de reconhecimento automático de fala – *Audimus* (Mendes et al., 2019). O *Audimus* é um sistema de reconhecimento de fala, baseado numa arquitetura modular, que recorre a um descodificador de fala baseado em Transdutores de Estado Finito (Weighted Finite-State Transducers - WFSTs), para combinar as probabilidades estimadas pelo modelo acústico, ao nível fonético, com aquelas resultantes do modelo de língua, ao nível da palavra, sendo o modulo de estimação das pronúncias das palavras a ponte entre esses dois módulos. As probabilidades fonéticas são estimadas recorrendo a uma abordagem híbrida, que usa as capacidades de modelação de sequências temporais dos Modelos de Markov Não Observáveis (Hidden Markov Models - HMMs) com as capacidades de identificação e segmentação de padrões das Redes Neuronais (Deep Neural Networks - DNNs). Podemos observar a arquitetura do sistema na Figura 1. Inicialmente, a fala é parametrizada em coeficientes acústicos, calculados a cada 10 ms (milissegundos). Em seguida, aplicando os modelos SNS (*Speech/Non-Speech*), a fala é segmentada. Desta forma, toda a fala humana é preservada e os segmentos que não são fala são filtrados, por exemplo eventos acústicos, ruídos e pausas. Após este processo, são utilizados modelos acústicos baseados em DNN-HMMs, para a realização da probabilidade dos fones das partes de fala do áudio. Nesta fase, o sistema estima as probabilidades de cada um dos fones condicionadas aos valores dos coeficientes acústicos de cada vector extraído anteriormente. Por exemplo, na primeira amostra da fala (t = 1), a onda sonora pode ter as seguintes probabilidades de fones - [e] - 80%, [g] - 10%, [i] - 10%. Utilizando os modelos acústicos, o sistema atribui a probabilidade fonética para descodificar o que falante pode ter dito foneticamente.

Posteriormente, utilizando a probabilidade fonética do passo anterior, o sistema descodifica os fones e as palavras através de modelos de pronúncia e de língua. Neste ponto, o sistema já possui a transcrição da fala humana, por exemplo, “>>[Ana] meu nome é ana”. O sinal >>, o exemplo anterior assinala a mudança de falante e [Ana] é o nome do falante que a deteção de tomada de palavra do falante (*Speaker Turn Detection*) e a identificação do falante (*Speaker Identification*) identificaram.

O resultado do módulo anterior ainda não é o produto final, porque é necessário grafar a maiúscula e pontuar a transcrição. O modelo de grafar as maiúsculas é treinado para grafar as primeiras letras das palavras em casos como a primeira letra do nome próprio ou a primeira letra de uma cidade, por exemplo. Os modelos de pontuação são treinados para pontuar as frases, de acordo com as regras gramaticais de cada língua. Assim, “>>[Ana] o meu nome é ana” é alterado para “>>[Ana] O meu nome é Ana”. O sistema da VoiceInteraction identifica três tipos de pontuação . , e ? , pelo que teria acrescentado o ponto no final do discurso – “>>[Ana] O meu nome é Ana.” O último módulo compreende a normalização do texto. Por exemplo, o que teria acontecido se o sistema de reconhecimento tivesse de compreender dígitos numéricos? Se este fosse o último passo, o sistema teria transcrito o dígito numérico 9 <nove> como “nove”, porque o vocabulário do modelo linguístico não possui numerais. Por isso, o sistema, no final de todos os processos de reconhecimento de fala, tem uma etapa de “Normalização do texto”, em que todos os dígitos são transcritos para dígitos numéricos – “nove” para 9.

Figura 1. Arquitetura do Sistema de Reconhecimento de Fala da *VoiceInteraction* - *Audimus* (apresentação oral de Sérgio Paulo no JEEC (2019))



Por último, o sistema é capaz de detetar e identificar as tomadas de palavra de cada falante. Assim, quando um falante chamado "Ana" está a falar, o sistema é capaz de identificar que é a Ana que está a falar e agrupar todas as suas falas do mesmo falante para ">>[Ana]...." (">>[Ana] O meu nome é Ana.").

5. Metodologia

Esta secção descreverá as etapas realizadas na análise da língua russa, que consistiu em quatro componentes principais:

i. A recolha de dados de fala e texto das diferentes regiões de língua russa;

ii. Revisão das transcrições de noticiários;

iii. Descrição do léxico russo previamente existente na *VoiceInteraction* e revisão das entradas mais frequentes do léxico, juntamente com os ajustes para pronúncias alternativas;

iv. Criação dos procedimentos metodológicos a utilizar na categorização das pausas preenchidas, encontradas ao longo das transcrições e não descritas anteriormente na literatura, tanto quanto nos é dado conhecer.

5.1. Recolha de dados

Foram recolhidos dados de fala e dados de texto de conteúdo noticioso das variedades da língua russa que continham reportagens, entrevistas de rua e debates políticos. O sistema utilizado para este trabalho já havia sido treinado com os dados de russo da Rússia europeia. Assim, foram recolhidos mais dados da variedade da Rússia europeia, incluindo agora também as variedades da Bielorrússia e do Cáucaso. Na Tabela 8, podemos observar os dados de fala recolhidos para cada variedade – 306 horas de fala extraídos de canais de televisão disponíveis em direto, de plataforma de *Youtube* e ainda de estações de rádio. Quanto aos dados de texto, foram recolhidos cerca de 500 milhões de palavras extraídos de jornais, artigos e corpora.

Tabela 8. Dados de fala recolhidos de Rússia europeia, Bielorrússia e Cáucaso

Regiões	**Dados de fala**
Rússia europeia	272h26m24s
Bielorrússia	24h44m40s
Cáucaso	8h48m59s
Total	306h00m03s

O **perfil dos falantes** dos dados recolhidos neste trabalho era constituído por: falantes nativos de língua russa de região de Rússia Europeia; falantes de língua segunda como no caso de falantes de Bielorrússia, visto que o russo faz parte das línguas oficiais do país; e falantes de língua franca provenientes do Cáucaso, já que o russo era falado na União Soviética e continua até hoje a ser utilizado pelos falantes.

5.2. **Transcrição dos dados**

Os dados de fala foram processados na plataforma proprietária, *Calligraphus*, na qual foram transcritos automaticamente e revistos por 4 anotadores humanos, falantes nativos de língua russa e com experiência na área. Para a realização de uma transcrição de qualidade, a *VoiceInteraction* criou um manual de anotação com todas as regras que os anotadores têm de seguir ao transcrever os dados de fala. Este manual já existia para as outras línguas, mas foi necessário verificar se o mesmo podia ser aplicado para a língua russa ou se eram necessárias algumas modificações ao mesmo. A título ilustrativo, algumas das regras do manual de anotação podem ser observadas na Tabela 9.

Tabela 9. Alguns dos símbolos utilizados pelos anotadores humanos no *Calligraphus*

Símbolo	**Uso**	**Exemplo**
[NOTRANS]	usado para indicar conteúdo impercetível ou conteúdo fora do domínio a transcrever (i.e., publicidade, música, entre outros.) que ocorra em grande parte ou na totalidade de um segmento de fala	*"Precisa de um aparelho auditivo? Ligue para o número apresentado no ecrã e aproveite a promoção de hoje."* > ***[NOTRANS]***
[UNK]	usado para indicar fragmentos de fala impercetíveis	*"Nós **[UNK]** escola."*
#	usado para indicar uma palavra soletrada	*"Enviei o ficheiro **#PDF**"*
_	usado como separador em palavras soletradas no plural	*"Ele mostrou os **#_P_D_F's**[a]."*
*	usado para indicar palavras truncadas	*"**Aman*** hoje nós vamos ao parque."*
+	used to indicate syncopated forms (for example, when the speakers pronounce "cause", the human annotator should transcribe it as "because" along with the symbol +)	*"We will go now cause it will rain tomorrow."* > *"We will go now **+because** it will rain tomorrow."*
%	usado para indicar pausas preenchidas	*"Eu **%@m** vou apresentar no ecrã."*
=	usado para indicar palavras prolongadas	*"Estas são **as=** novidades de hoje."*

[a] Se o underscore não fosse adicionado, o "s" seria reconhecido pelo sistema como sendo uma letra soletrada > [pedeɛfɛs] ao invés de [pedeɛfʃ].

Os anotadores transcreveram o conteúdo do áudio de acordo com os símbolos apresentados na Tabela 9. Para garantir que o reconhecedor de fala só é treinado com dados de fala de qualidade e que a transcrição é o

mais próxima possível do que foi proferido, também foi verificado se os anotadores seguiram o manual de anotação. Foram utilizadas 4 horas de áudio (um áudio de cada canal de televisão) como amostra para rever o manual de anotação e observar como os anotadores humanos o utilizam, analisar o desempenho do reconhecedor de fala nas diferentes variedades da língua russa e, finalmente, observar o desempenho dos anotadores humanos.

A verificação das transcrições teve como objetivo verificar os seguintes aspetos:

i. verificar se o manual de anotação era replicável para a língua russa;

ii. determinar quais seriam as áreas mais problemáticas para transcrever automaticamente e o seu impacto nos anotadores humanos;

iii. verificar se os anotadores humanos seguiram o manual de anotação;

iv. analisar as abordagens baseadas nos dados recolhidos para melhorar o léxico com a potencial adição de novas palavras e novas pronúncias para palavras já muito frequentes;

v. analisar o conjunto de pausas preenchidas para a língua russa no *Audimus*, com base na literatura.

5.3. **Análise das transcrições**

Analisámos as transcrições das diferentes regiões falantes de língua russa, utilizando o manual de anotação criado para as outras línguas. Assim, foi possível estabelecer as estruturas problemáticas para um reconhecedor de fala transcrever e, subsequentemente, o seu impacto nos transcritores. Áreas como a sobreposição de fala, o ruído de fundo e as disfluências, entre outros mecanismos de fala e eventos acústicos, demonstraram ser complexas de anotar para o reconhecedor de fala e para os anotadores humanos. Deste modo, prosseguimos com a resolução dessas estruturas.

De forma a resolver alguns dos erros mais frequentemente ocorridos nas transcrições, foi criada uma lista dos erros e a sua respetiva resolução. O objetivo é transmitir esta informação aos transcritores, para melhorar o seu desempenho nas transcrições da língua russa e incluí-la no futuro manual de anotação.

Um dos erros mais comuns foi o de não colocação/sinalização das pausas preenchidas. O sistema ASR não reconheceu as pausas preenchidas e os anotadores humanos também não as identificaram ou transcreveram-nas como discurso imperceptível - [UNK] / [NOTRANS]. Podemos observar um exemplo na Figura 2, antes da correção dos erros, e na Figura 3, após a correção.

Figura 2. Erros antes da correção

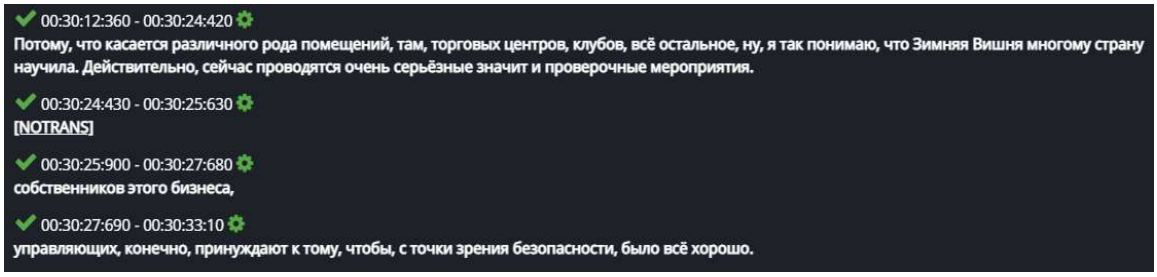
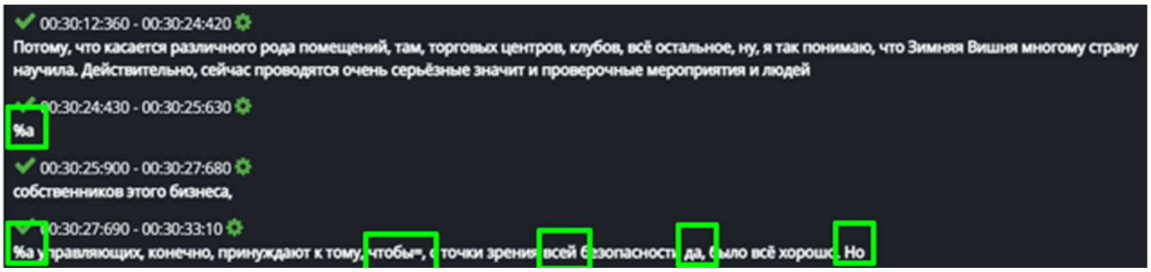


Figura 3. Erros após a correção



Na Figura 2, observamos no minuto 00:30:24:430 que a pausa preenchida foi transcrita como um conteúdo impercetível - [NOTRANS] e no minuto 00:30:27:690 a pausa preenchida - %a, no início do segmento de fala, foi ignorada e não transcrita. Na Figura 3, procedeu-se à correção de todos os erros, assinalados com quadrados verdes.

Para além das pausas preenchidas em falta, podemos observar na Figura 3, no minuto 30:27:690 a palavra "чтобы" 'para' apresenta agora um símbolo "=", indicando que foi pronunciada com um prolongamento. Além disso, foram acrescentadas palavras no mesmo minuto – "всей" 'inteira', "да" 'sim' e "Но" 'mas' - uma vez que nem o sistema, nem o anotador as transcreveu.

As palavras repetidas, interjeições e palavras truncadas também fizeram parte das estruturas problemáticas para os anotadores. Os anotadores humanos tentaram obter uma transcrição "bem estruturada" e por isso, ignoraram as disfluências e os marcadores discursivos para obter as frases gramaticalmente corretas. Também foram encontrados e corrigidos erros como a falta de grafar a maiúscula, normalização ou pontuação. No futuro, será criado o manual de anotação para a língua russa com o mesmo tipo de erros mais comuns, com o objetivo de melhorar o desempenho dos anotadores humanos e do reconhecedor de fala. Até lá, esta informação será transmitida ao departamento responsável pelo contacto com os anotadores da *VoiceInteraction*, para que estes recebam *feedback* e possam aplicá-lo.

As transcrições foram avaliadas utilizado o WER (*Word Error Rate*) que foi calculado através da Equação 1.

Equação 1. Equação usada para calcular o WER

$$WER = \frac{SUBS + DEL + INS}{N}$$

Quanto mais baixo for o WER, melhor a precisão do sistema de reconhecimento. Por exemplo, um WER de 10% significa que 90% da transcrição encontra-se correta. Na equação apresentada acima, SUBS é o número de substituições que foram efetuadas no texto de *output* em comparação com o que foi reconhecido automaticamente, DEL é o número de apagamentos e INS é o número de inserções. Finalmente, N é o número total de palavras de uma transcrição.

5.4. **Léxico**

A *VoiceInteraction* dispunha de um léxico previamente criado para a língua russa, baseado em regras fonéticas da língua russa. O número de entradas do léxico era superior a 400 mil entradas. Estas foram organizadas a partir das mais frequentes nos dados de fala para as menos frequentes. De 409447 pronúncias, 10 mil palavras mais frequentes foram revistas e modificadas, para permitir a extração de padrões a utilizar no módulo estatístico de G2P (*Grapheme-to-phone*). As pronúncias mais frequentes foram usadas como amostra, para observar se as regularidades e regras fonéticas foram seguidas e se as pronúncias eram as pronúncias mais frequentes e fiáveis para cada palavra.

Foi possível detetar erros comuns nas 10 mil pronúncias mais frequentes, como se exemplifica:

— letra "e" em posição final da palavra tinha, em alguns casos, uma pronúncia de [ə] em vez de [e], como em *основное* [ʊsnʌvnójə] > [ʊsnʌvnój**e**] 'principal';

— palavras com a pronúncia de [ɪ] em posição tónica em vez de [i], como em *мило* [mʲílə] > [mʲ**í**ɬə] 'agradável' e vice-versa;

— em algumas pronúncias a partícula palatalizadora [j] antes dos fonemas [ɪ] e [i] encontrava-se em falta, como em *часы* [**t͡ɕ**ɪsɨ́] > [**t͡ɕ**ʲɪsɨ́] 'relógio', por exemplo.

Também foram encontrados erros que ocorreram de forma aleatória, apenas uma ou duas vezes na amostra analisada. Por exemplo, a letra "в", em vez de ter pronúncias correspondentes de [v] / [f] / [vɛ], tinha também uma pronúncia de [vʌstók]. Neste caso, verificamos se esta pronúncia particular - [vʌstók] - estava no léxico associado à palavra correspondente - *восток* 'Este'. A verificação foi realizada com recurso a uma função de linha de comando chamada "grep". Após a verificação, a correspondência entre "в" e [vʌstók] foi eliminada por estar incorreta.

Além disso, foram incluídas pronúncias alternativas de acordo com as variedades da língua russa. Por exemplo, a partícula palatalizadora, de acordo com a variedade, é realizada ou não, como em *следовательно* [**s**ʲlʲédəvətʲɪlʲnə] 'portanto', a primeira consoante [s] é pronunciada com palatalização, ou [**s**lʲédəvətʲɪlʲnə] em que o [s] é pronunciado sem palatalização. A língua russa também apresenta palavras homógrafas. Normalmente, o léxico continha apenas um dos significados. O significado e a pronúncia alternativos foram acrescentados. Por exemplo, *пропасть*, se pronunciado como [**pró**pəstʲ] com o acento na primeira sílaba, tem o significado de 'abismo'. No entanto, se a palavra for pronunciada com o acento na segunda sílaba, como em [prʌ**pástʲ**], significa 'desaparecer'.

Para contagem de número de substituições e modificações ao léxico, foi utilizada uma linha de comando chamada "grep". A expressão regular utilizada para comparar os dois léxicos, antes e depois das modificações, é apresentada na Figura 4:

Figura 4. Expressão regular utilizada para comparar os dois léxicos

```
fgrep -vf lexicon-antes.txt lexicon-depois.txt
```

5.5. **Pausas preenchidas**

A análise das transcrições permitiu identificar as pausas preenchidas mais frequentemente utilizadas pelos falantes de russo, já descritas na literatura e indicadas na Tabela 7.

Foi realizada uma análise acústica e espectrográfica de cada uma das pausas preenchidas. Estas foram extraídas manualmente da amostra dos áudios, utilizando o *software Praat* (Boersma & Weenink, 1999–2022) e guardadas em formato *.wav* para o treino do reconhecedor.

Após a extração de todas as pausas preenchidas, cada pausa foi organizada pela sua duração, de mais curta para a mais longa. Desta forma, o reconhecedor de fala foi treinado com diferentes durações de cada pausa para as reconhecer automaticamente. Além disso, cada pausa foi adicionada ao léxico de língua russa, apresentando pronúncias distintas de acordo com a duração de cada pausa. Por exemplo, uma pausa preenchida *%a* poderia ser pronunciada como [a], mas também como [aaa], com uma duração maior, ou *%ам* que poderia ser pronunciada como [am] e [aam], entre outras variações.

Na literatura, a maioria das pausas preenchidas encontradas já havia sido estudada, com exceção de duas que, até onde sabemos, não haviam sido descritas. Para categorizar as duas pausas preenchidas restantes, utilizámos o *software Praat*, que permitiu extrair o espectrograma e observar a frequência, os formantes, a duração e a intensidade de cada pausa preenchida. Estas serão apresentadas na secção de Resultados.

6. **Resultados**

O presente capítulo irá descrever os resultados da análise das transcrições; os resultados antes e depois das modificações realizadas às 10 mil entradas mais frequentes do léxico; as pausas preenchidas que foram encontradas ao longo das transcrições, juntamente com a sua frequência nas transcrições analisadas; e aplicação do presente trabalho à um projeto Europeu AIDA (Artificial Intelligence and Advanced Data Analysis for Authority Agencies).

6.1. **Análise das transcrições**

Através da análise das transcrições, foi possível verificar que o manual de anotação já existente para as outras línguas também era replicável para a língua russa. Os anotadores humanos seguiram o manual de anotação de outras línguas para transcrever a língua russa, visto que o mesmo não havia sido validado anteriormente.

Por conseguinte, estabelecemos as estruturas problemáticas para transcrever automaticamente e o seu impacto nos transcritores humanos: fala sobreposta, pausas preenchidas, palavras prolongadas, repetições, palavras truncadas, palavras soletradas, pontuação e grafia da letra maiúscula. Para a obtenção de um discurso "bem estruturado", os anotadores humanos eliminaram todos os mecanismos de fala que os falantes utilizam, criando enunciados longe da realidade e não seguindo o manual de anotação fornecido pela *VoiceInteraction*.

Na Tabela 10, podemos observar os resultados de comparação das transcrições automáticas *versus* as transcrições manuais.

Na primeira coluna encontra-se a divisão em três regiões de língua russa: Russo europeu (com cinco ficheiros correspondentes das diferentes campanhas), a região do Cáucaso e a Bielorrússia, com um ficheiro cada. Cada ficheiro apresenta o WER antes e depois da validação das transcrições e a correspondente diferença.

Tabela 10. Comparação das transcrições automáticas *versus* as transcrições manuais

	Antes	**Depois**	**Diferença (%)**
Variedade	**WER (%)**	**WER (%)**	
Rússia Europeia	18,32	18,17	-0,15
	6,86	6,88	0,02
	16,87	16,22	-0,65
	10,67	9,24	-1,43
	12,39	12,48	0,09
Cáucaso	**24,43**	**22,71**	**-1,71**
Bielorrússia	**20,57**	**18,69**	**-1,88**

Nota. A negrito encontram-se os valores de WER que apresentam a maior descida após a validação das transcrições automáticas.

Na maioria dos casos, observamos a diminuição do WER, como esperado. No entanto, também observámos um aumento do WER em dois casos, 0,02% e 0,09%. Embora tenhamos corrigido as transcrições, os modelos não foram suficientemente eficazes para identificar as palavras corretas, potencialmente problemáticas. Pode ser devido aos erros do sistema que estamos a eliminar e as novas alterações são detetadas como incorretas. Além disso, os áudios com sobreposição de fala, disfluências e eventos acústicos influenciam o WER.

Para o russo europeu, o WER mais elevado foi de 18,32%, antes da validação, e de 18,17%, após a validação. A nossa explicação para um WER tão elevado é a presença de debate político no áudio com fala sobreposta, áudio de má qualidade e disfluências, o que resulta em problemas de ASR.

Quanto as variedades do russo do Cáucaso e da Bielorrússia, o WER é superior ao da variedade da Rússia europeia - 24,43% e 20,57%. Uma vez que o ASR só foi treinado com a variedade do russo europeu, foi a primeira vez que teve de reconhecer outras variedades. Após a validação, o WER diminuiu em ambos os casos, 1,71% para a região do Cáucaso e 1,88% para a Bielorrússia. Para diminuir ainda mais o WER, seria necessário ajustar o ASR a cada variedade, por exemplo, ajustar as pronúncias, os modelos acústicos, entre outros, assim como incluir mais dados.

6.2. **Léxico**

O léxico apresentado pela *VoiceInteraction* continha 400 mil entradas, a partir das quais foram revistas e validadas 10 mil palavras mais frequentemente observadas num corpus de texto recolhido de *websites* de jornais de referência russos. A verificação do léxico, juntamente com a adição de novas pronúncias, baseou-se no estado da arte da língua russa, tal como descrito anteriormente. Após a validação do léxico, foram incluídas 145 novas entradas e 86 foram modificadas, totalizando 231 alterações. O léxico recentemente revisto, composto por 10 145 entradas, foi utilizado para avaliar a ASR após as modificações.

Foram efetuadas duas avaliações: antes e depois das modificações do léxico. Os resultados preliminares sem retreino dos modelos acústicos não demostram resultados significativos - 19,85% de WER antes das modificações e 19,97% de WER depois, com uma diferença de 0,12%. Os resultados não refletem imediatamente a melhoria devido a um vocabulário extenso de 400 mil entradas. Além disso, novos erros passaram a ser produzidos pelo ASR. Por exemplo, palavras anteriormente reconhecidas pelo reconhecedor não foram reconhecidas após as modificações efetuadas - a palavra *наверное* 'provavelmente' foi anteriormente reconhecida corretamente, mas após as modificações, foi reconhecida como *наверно* com o último grafema *е* em falta. No entanto, foi possível observar um melhor desempenho do sistema de reconhecimento em algumas pronúncias, corrigindo erros anteriormente produzidos. Algumas dessas palavras eram palavras funcionais – *в*, *и*, *за*, acrónimos – *ВМФ*, adjetivos – *главная*, verbos – *ловлю* e substantivos – *курение*, *семь*.

Embora os resultados preliminares não tenham demonstrado resultados significativos, pudemos proceder à validação do léxico com base na literatura e criar e incluir novas regras fonéticas, descritas na secção de "Metodologia":

- o grafema *е* em posição final de palavra com a pronúncia de [e] (em vez de [ə]);
- [i] sempre em posição tónica (em vez de [ɪ]);
- partícula palatalizadora [j] antes dos fonemas [ɪ] e [i];
- [ɛ] em início de palavra, quando em posição átona com as correspondentes pronúncias de [ɛ], [ɪ] e [ɨ].

Foi notório que as regras fonéticas e o conjunto de fones criados para a língua russa na *VoiceInteraction* seguiram a literatura sobre a fonética da língua russa. É essencial realçar a importância da validação do vocabulário pois, dessa forma, podemos ter um modelo de pronúncia e um modelo de léxico bem construídos,

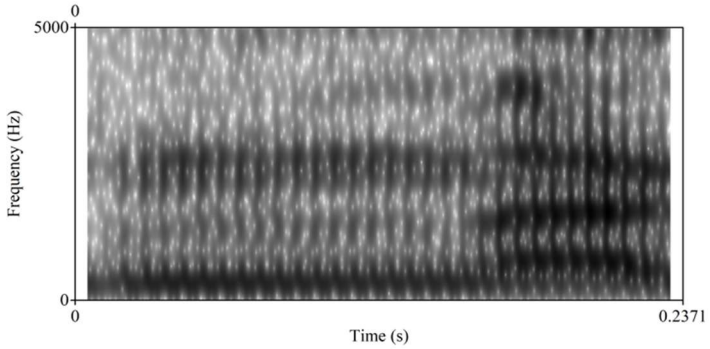
produzindo transcrições no futuro o mais próximo possível das produções reais. A qualidade do léxico não foi previamente verificada na *VoiceInteraction*. Assim, ao revermos o léxico e ao estabelecermos a criação de novas regras fonéticas a serem alargadas às 400 mil entradas garantimos a eliminação de erros sistemáticos.

6.3. **Pausas preenchidas**

Dados de fala foram recolhidos de noticiários transmitidos na Rússia europeia, Bielorrússia e na região do Cáucaso. Os dados recolhidos foram transcritos utilizando a plataforma *Calligraphus*, anotando os mecanismos de fala e as disfluências o mais próximo possível das produções reais.

Após uma análise das transcrições, identificámos duas pausas preenchidas adicionais que, tanto quanto sabemos, ainda não foram descritas na literatura - *%на* e *%мна*. Extraímos o espectrograma de cada uma destas pausas preenchidas encontradas ao longo da análise da transcrição. A primeira - *%на*, é pronunciada como [na], e o seu espectrograma é apresentado na Figura 5.

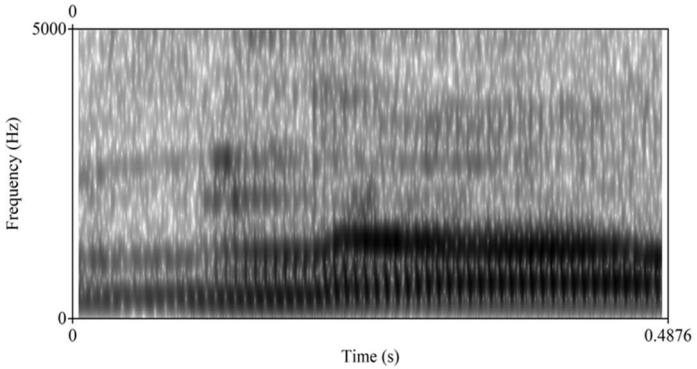
Figura 5. Espectrograma da pausa preenchida *%на*



Na Figura 4, é visível através de energia menos intensa a consoante [n] e através de formantes mais intensos, no final, é visível a vogal [a]. Para a nasal alveolar, [n], tal como apresentado em Ladefoged (2003), o F1 tende a ser baixo - 250-300 Hz e o F2 a rondar os 2500 Hz. Entre os formantes, como podemos observar, também há pouca energia (onde o som [a] tem sua F2). A vogal [a] para o F1 apresentou valores entre 640-720 Hz e 1200-1450 Hz para o F2.

Ao mesmo tempo, observamos a pausa preenchida *%мна,* pronunciada como [mna], apresentada na Figura 6.

Figura 6. Espectrograma da pausa preenchida *%мна*



No início do espetrograma, observamos um segmento com menor intensidade, que é a consoante [m], após esse segmento com maior intensidade é apresentada a consoante [n], e o último segmento apresenta visualmente maior energia, indicando que estamos na presença da vogal [a], também com maior duração. Assim como o som [n], o som [m] apresentou valores de F1 entre 250-300 Hz e F2 em torno de 2500 Hz, com energia entre os dois formantes.

A Tabela 11 apresenta a distribuição das pausas preenchidas em cada ficheiro analisado, juntamente com a distribuição das pausas preenchidas *%на* e *%мна*.

Tabela 11. Distribuição das pausas preenchidas nos ficheiros analisados

Variedade	**Pausas preenchidas (%)**	***%на* (%)**	***%мна* (%)**
Rússia Europeia	0,7	0,04	0
	0,4	0	0
	1,3	0,26	0,05
	3,5	0,31	0,1
	2,7	0,02	0
Cáucaso	**3,3**	**0**	**0**
Bielorrússia	0,7	**0**	**0**

Nota. Na segunda coluna, a negrito encontram-se os ficheiros com maior taxa de frequência das pausas preenchidas observadas; nas colunas terceira e quarta, encontram-se a negrito os valores das variedades nas quais não foram observadas pausas preenchidas %на e %мна.

Na variedade do russo europeu, os ficheiros com a maior frequência de pausas preenchidas apresentam valores de 3,5% e 2,7%. A explicação para este fenómeno deve-se ao tipo de conteúdo dos ficheiros - debate político. Uma vez que o discurso apresentado não foi planeado e foi muito emotivo, favoreceu a produção das pausas preenchidas e, posteriormente, contribuiu para a produção de outros tipos de disfluências, como repetições, truncamentos (devido à fala sobreposta), prolongamentos, erros de pronúncia, entre outros.

Na região do Cáucaso, o ficheiro analisado também apresentou uma alta frequência das pausas preenchidas - 3,3%. Neste caso, o discurso foi maioritariamente proveniente das entrevistas de rua, correspondendo a um discurso espontâneo repleto de diversas disfluências.

De acordo com os resultados, as pausas preenchidas *%на* e *%мна* só foram encontradas na variedade do russo europeu. A produção de *%на* foi mais frequente do que a de *%мна*. No entanto, estas foram produzidas apenas por falantes nativos, como políticos, repórteres, entrevistadores e jornalistas. Na região do Cáucaso e na Bielorrússia, estas pausas preenchidas não foram encontradas. Os resultados podem sugerir que *%на* e *%мна* estão a ser produzidas apenas por falantes nativos de russo. No entanto, para confirmar essa hipótese, seria necessário alargar o conjunto de dados para confirmar a hipótese e analisar outras variedades de russo.

7. **Projeto AIDA**

O projeto *Artificial Intelligence and Advanced Data Analysis for Authority Agencies* (AIDA) é um projeto europeu que visa detetar possível *cybercrime* e terrorismo. A *VoiceInteraction*, enquanto membro do projeto, contribuiu com a transcrição de dados de várias línguas, entre elas o russo, e outras informações estritamente confidenciais. Uma vez que os dados de fala eram conteúdos sensíveis fornecidos pela polícia e confidenciais, a nova versão do *Calligraphus* foi criada especialmente para o AIDA, apenas disponível para as pessoas que trabalham diretamente no projeto.

Os parceiros do projeto forneceram-nos dados no domínio da propaganda terrorista, naturalmente gravados em condições acústicas muito difíceis para as tecnologias de transcrição automática. No total, foram recolhidas 7 horas de dados de fala, e cerca de 3 horas foram transcritas e revistas. Foi um desafio transcrever estes áudios, uma vez que os oradores eram falantes nativos de árabe e por isso possuíam um sotaque forte na língua russa: as palavras tinham pronúncias diferentes do russo europeu; a concordância entre verbo e sujeito, na maioria dos casos, estava ausente; e os oradores usavam frequentemente termos, nomes ou conceitos na sua língua materna – árabe, desconhecidos para nós e inexistentes no vocabulário do sistema.

Como mencionado, foram transcritas cerca de 3 horas de fala e ocorreram 3113 substituições devido a: palavras inexistentes no vocabulário do sistema, por exemplo, *Аллах* 'Alá'; palavras incorretamente reconhecidas; e discurso estrangeiro que não é transcrito, marcado com [NOTRANS], mas que o ASR transcreveu incorretamente como sendo da língua russa.

Tabela 12. Palavras mais frequentemente substituídas nos ficheiros transcritos

Palavra	**SUBS**	**SUBS (%)**
Аллаха	112	3,6
Аллах	100	3,2
И	85	2,7
На	23	0,7
Не	22	0,7
Кафиров	20	0,6
В	20	0,6
По	17	0,5
Мы	16	0,5
И	16	0,5

As substituições mais frequentes (cf. Tabela 12) ocorreram em palavras relacionadas com a religião e em palavras funcionais. Uma vez que o modelo linguístico foi adaptado para o conteúdo noticioso, as palavras de conteúdo religioso não foram reconhecidas. As palavras funcionais também fazem parte das substituições mais frequentes, na maioria dos casos, devido à coarticulação com a palavra prévia ou adjacente. Além disso, as palavras funcionais têm uma duração curta e podem ser absorvidas pelos modelos acústicos durante a descodificação da fala, resultando numa transcrição incorreta.

8. **Conclusões**

Inicialmente, validámos o manual de anotação a ser aplicado ao russo. Através da análise das transcrições de material noticioso, pudemos confirmar as estruturas mais problemáticas tanto para o ASR como para os anotadores humanos, tais como disfluências, fala sobreposta e palavras funcionais. Em trabalhos futuros, tencionamos adaptar o manual de anotação para a língua russa que abordem as estruturas problemáticas e fornecer a sua resolução.

Os áudios com uma maior ocorrência de fala sobreposta e diferentes disfluências, como pausas preenchidas, prolongamentos e truncamentos, apresentaram um WER mais elevado - para o russo europeu o WER mais elevado foi de 18,32% antes da validação e de 18,17% após a validação. Observámos um aumento do WER após as modificações em dois ficheiros - 0,02% e 0,09%. Apesar das nossas modificações nas transcrições, foi difícil para os modelos identificarem as palavras corretas, potencialmente problemáticas. Além disso, o áudio com fala sobreposta, disfluências e eventos acústicos influenciam o WER. As variedades da região do Cáucaso e da Bielorrússia também apresentaram um WER mais elevado - 24,43% e 20,57%, porque o ASR só foi treinado com a variedade do russo europeu. Após a validação das transcrições, foi possível

diminuir o WER em ambas as regiões - 1,71% para a região do Cáucaso e 1,88% para a Bielorrússia. No futuro, o nosso objetivo é ajustar a ASR a cada variedade de língua russa, por exemplo, ajustar as pronúncias, os modelos acústicos, entre outros.

Foi confirmado que o léxico analisado das 10 mil palavras mais frequentes estava de acordo com as regularidades e regras fonéticas russas, exceto em alguns casos, como o grafema "e" em posição final da palavra com a pronúncia de [ə] (em vez de [e]) e [I] em posição tónica (em vez de [i]). No futuro, esperamos alargar as correções às 400 mil entradas, de modo a garantir a eliminação de erros sistemáticos e a obter um modelo de pronúncia e um modelo léxico bem construídos, produzindo transcrições tão próximas quanto possível das produções reais.

Embora os resultados preliminares sem retreino dos modelos acústicos não mostrem resultados significativos - 19,85% WER antes das modificações e 19,97% WER depois, com uma diferença de 0,12%, observámos uma melhoria em algumas das palavras corrigidas, tais como palavras funcionais, acrónimos, adjetivos, verbos e substantivos. No futuro, deve ser efetuada uma análise do reconhecimento de erros para diminuir o WER, como a análise dos erros, uma análise lexical mais alargada e uma análise dos modelos acústicos russos.

Além disso, validámos o conjunto de pausas preenchidas e acrescentámos duas ainda não descritas na literatura, tanto quanto sabemos, através da análise das transcrições - *%на* e *%мна*. De acordo com os resultados, estas só foram produzidas na variedade do russo europeu - *%на* com uma frequência de 0,63% e *%мна* com 0,15%. Os falantes não produziram estas pausas preenchidas na Bielorrússia e na região do Cáucaso. Assim, os resultados sugerem que estas pausas podem estar a ser produzidas apenas por falantes nativos. No entanto, seria necessário analisar um conjunto alargado de dados da Bielorrússia e da região do Cáucaso, juntamente com as outras variedades de língua russa, para confirmar a hipótese.

O trabalho realizado foi aplicado a um projeto europeu – AIDA. Através da validação do manual de anotação, da validação das transcrições, da validação das regras e regularidades fonéticas e da validação das pausas preenchidas, foi possível aplicar o trabalho ao projeto de transcrição da fala. Os dados deste projeto eram difíceis de reconhecer devido ao forte sotaque dos falantes na língua russa, os falantes não seguiam as regras gramaticais - por exemplo, a concordância entre verbo e sujeito encontrava-se em falta e os falantes usavam frequentemente termos, nomes ou conceitos na sua língua materna que não reconhecíamos.

Após a análise das transcrições, pudemos observar um elevado número de substituições. O modelo linguístico foi adaptado ao conteúdo noticioso. Por conseguinte, muitas das palavras russas foram incorretamente reconhecidas ou não foram reconhecidas de todo, uma vez que não faziam parte do vocabulário. As palavras funcionais também foram incorretamente reconhecidas, na maioria dos casos, devido à coarticulação com a palavra anterior ou seguinte, ou devido à sua curta duração - as palavras funcionais podem ser absorvidas pelos modelos acústicos durante a descodificação da fala.

Por último, a validação de sistema ASR da língua russa já contribuiu para o projeto europeu AIDA e ainda para *Young Female Researchers in Speech Workshop* (2022), um evento satélite de *Interspeech*, realizado em Incheon, Coreia do Sul. O presente trabalho reflete a identificação de duas novas pausas preenchidas, informação crucial para os sistemas ASR e para a diminuição do WER. A metodologia utilizada neste trabalho juntamente com os procedimentos, poderão ser replicados para a criação e validação de outros sistemas ASR.

Referências

Adell, Jordi, David Escudero & Antonio Bonafonte (2012) Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. *Speech Communication* 54 (3), pp. 459–476. https://doi.org/10.1016/j.specom.2011.10.010

Allwood, Jens, Joakim Nivre & Elisabeth Ahlsèn (1990) Speech management: On the non-written life of speech. *Nordic Journal of Linguistics* 13 (1), pp. 3–48. https://doi.org/10.1017/S0332586500002092

Arnold, Arnold, Maria Fagnano & Michael Tanenhaus (2003) Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research* 32 (1), pp. 25–36. https://doi.org/10.1023/a:1021980931292
Barczewska, Katarzyna & Magdalena Igras (2012) Detection of disfluencies in speech signal. *Challenges of modern technology*, 4 (2), pp. 3–10.
Bear, John, John Dowding & Elizabeth Shriberg (1992) Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 56–63.
Betz, Simon (2020) *Hesitations in spoken dialogue systems*. Tese de doutoramento, Universitat Bielefeld.
Boersma, Paul & David Weenink (1992–2022) *Praat: Doing phonetics by computer* (Versão 6.2.14) [Programa de computador]. Disponível em https://www.praat.org
Bogdanova-Beglarian, Natalia & Ekaterina Baeva (2018) Nonverbal elements in everyday Russian speech: An attempt at categorization. In *Proceedings of Computational Models in Language and Speech Workshop (CMLS 2018)*. Akademija nauk Respubliki Tatarstan, pp. 3–13. Disponível em http://ceur-ws.org/Vol-2303/
Borunova, Svetlana, Vera Vorontsova & Natalia Es'kova (1989) Orfoepicheskii slovar' russkogo yazyka. Proiznoshenie. Udarenie. Grammaticheskie formy [Dicionário ortoépico da língua russa. Pronúncia. Ênfase. Formas gramaticais] (5.ª ed.). Russkii yazyk Publ.
Butzberger, John, Hy Murveit, Elizabeth Shriberg & Patti Price (1992) Spontaneous speech effects in large vocabulary speech recognition applications. In *Proceedings of the 1992 DARPA Speech and Natural Language Workshop*. Morgan Kaufmann, pp. 339–343.
Clark, Herbert & Jean Fox Tree (2002) Using uh and um in spontaneous speaking. *Cognition* (84), pp. 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3
Demidov, Dmitriy, Vasiliy Klepatskiy, Elena Morozova, Michail Popov, Dmitriy Rudnev, Anastasia Ryko, Dmitriy Cherdakov & Olga Cherepanova (2013) *Istoricheskaya grammatika russkogo yazyka: Khrestomatiya* [Gramática histórica da língua russa: Um livro didático]. Izdatel'stvo Sankt-Peterburgskogo universiteta.
Gabrea, Marcel & Douglas O'Shaughnessy (2000) Detection of filled pauses in spontaneous conversational speech. In *Proceedings Sixth International Conference on Spoken Language Processing (ICSLP 2000)* (Vol. 3), pp. 678–681. https://doi.org/10.21437/ICSLP.2000-626
Kharlamova, Tatiana L. (2008). Sopostavitelnyj analiz zvukov zapolnitelej pauz hezitacii, kak vyrazitelej chuvstvovanij, v russkom i anglijskom yazykah. [Análise comparativa dos sons das pausas preenchidas em russo e inglês, como expressores de sentimentos].
Heeman, Peter & James Allen (1994) Detecting and correcting speech repairs. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 295–302. Disponível em https://aclanthology.org/P94-1041
Heeman, Peter A. & Allen, James F. (2001). Improving robustness by modeling spontaneous speech events. In *Robustness in language and speech technology* (pp. 123-152). Dordrecht: Springer Netherlands.
Hough, Julian & Matthew Purver (2012) Processing self-repairs in an incremental type-theoretic dialogue system. In *Proceedings of the SemDial 2012 (SeineDial): The 16th Workshop on the Semantics and Pragmatics of Dialogue*. SemDial, pp. 136–144. Disponível em https://www.semdial.org/anthology/papers/Z/Z12/Z12-3018/
Jurafsky, Dan & James Martin (2021). *Speech and language processing*. [Manuscrito em preparação]
Kasatkin, Leonid (2014) *Sovremennyy russkiy yazyk. Fonetika* [Língua russa moderna. Fonética]. (3.ª ed.). Knizhnyi dom 'LIBROKOM'.
Levelt, Willem (1983) Monitoring and self-repair in speech. *Cognition* (14), pp. 41–104.
Litnevskaya, Elena (2006) *Russkiy yazyk: kratkiy teoreticheskiy kurs dlya shkol'nikov* [Língua russa: Um breve curso teórico para estudantes]. Disponível em http://www.gramota.ru/
Liu, Yang, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf & Mary Harper (2006) Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE*

Transactions on audio, speech, and language processing, 14 (5), pp. 1526–1540. https://doi.org/10.1109/TASL.2006.878255
Mendes, Carlos, Alberto Abad, João Neto & Isabel Trancoso (2019) Recognition of Latin American Spanish Using Multi-Task Learning. In *Proceedings INTERSPEECH 2019*. ISCA, pp. 2135–2139. Disponível em https://www.isca-speech.org/archive/interspeech_2019/mendes19_interspeech.html
Moniz, Helena (2013) *Processing disfluencies in European Portuguese*. Tese de doutoramento, Universidade de Lisboa.
Nakatani, Christine & Julia Hirschberg (1994) A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America* 95 (3), pp. 1603–1616. https://doi.org/10.1121/1.408547
Osipova, Tamara (2018) *Fonetika sovremennogo russkogo yazyka* [Fonética da língua russa moderna]. Knizhnyy dom "LIBROKOM".
Oviatt, Sharon (1995) Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language* 9 (1), pp. 19–35. https://doi.org/10.1006/csla.1995.0002
Paulo, Sérgio (2019, 15–11 março) *Automatic subtitling: Pushing the limits of speech recognition* [Apresentação de comunicação]. Engineering and Tech Talks - JEEC 2019, Instituto Superior Técnico, Lisboa, Portugal.
Popov, Michail (2014) *Fonetika sovremennogo russkogo yazyka* [Fonética da língua russa moderna]. Filologicheskiy fakul'tet Sankt-Peterburgskogo gosudarstvennogo universiteta.
Schettino, Loredana (2022) *The role of disfluencies in Italian discourse. Modelling and speech synthesis applications*. Tese de doutoramento, Universita Degli Studi di Salerno.
Shriberg, Elizabeth (2001) To ‘errrr’is human: ecology and acoustics of speech disfluencies. *Journal of the international phonetic association* 31 (1), pp. 153–169. https://doi.org/10.1017/S0025100301001128
Sokolova, Anastasia (2021) *Fonetika i fonologiya russkogo yazyka* [Fonética e fonologia da língua russa]. Masarykova Univerzita.
Swerts, Marc (1998) Filled pauses as markers of discourse structure. *Journal of Pragmatics* (30), pp. 485–496. https://doi.org/10.1016/S0378-2166(98)00014-9
Teng, Hai & Svetlana Androsova (2022) Osobennosti pauzacii v rodnoj i akcentnoj spontannoj rechi: akusticheskij analiz kitajskoj rodnoj, russkoj akcentnoj i russkoj rodnoj rechi [Características das pausas na fala espontânea nativa e com sotaque: uma análise acústica da fala nativa chinesa, da fala com sotaque russa e da fala nativa russa]. *Vestnik NGU* 21 (2), pp. 67–86.
Teng, Hai (2015) Universalnye i tipologicheskie cherty pauzacii v spontannoj rechi nositelej raznyx yazykov. [Características universais e típicas das pausas na fala espontânea de falantes de diferentes línguas]. *Teoreticheskaya i prikladnaya lingvistika* [Linguística teórica e aplicada] 1 (2), pp. 105–113.
Yanushevskaya, Irena & Daniel Bunčić (2015) Russian. *Journal of the International Phonetic Association* 45 (2), pp. 221–228. https://doi,org/10.1017/S0025100314000395
Yearley, Jennifer (1995) Jer vowels in Russian. In *University of Massachusetts occasional papers in linguistics* (Vol. 18). University of Massachusetts, pp. 533–571.