Tailoring Furhat robotic head lip-syncing to Galician language: an adaptation and evaluation study

Carla Castedo[1,*], Carmen Magariños[2,3], Alejandro Catala[1,3] and Alberto Bugarín-Diz[1,3]

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Spain
[2]Instituto da Lingua Galega, Universidade de Santiago de Compostela, Spain
[3]Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, Spain

Abstract

Text-speech alignment and lip-syncing are crucial for a pleasant interaction with an embodied conversational agent, especially for anthropomorphic social robots like Furhat. While pre-trained alignment models are integrated into these agents, they may not fully align the target language, as they are primarily trained in English. This paper addresses this limitation for Galician, leveraging a previously developed text-to-speech system, the Furhat robot, and the Montreal Forced Aligner (MFA). We create acoustic models and a pronunciation dictionary for Galician from scratch, which is a key contribution given the lack of resources. We propose an alternative method using MFA to generate accurate phone-level alignments for Galician synthetic speech and evaluate its quality through objective and subjective experiments. In this preliminary study, our trained model's accuracy in misalignment assessment matches results reported in the literature for other languages despite the limited data availability for Galician. Regarding the subjective evaluation, a perceptual test with native speakers reveals a strong preference (88%) for our lip synchronization over Furhat's default (2%), highlighting the validity of our method for improving lip-syncing in under-resourced languages.

Keywords

lip-syncing, forced alignment, social robot, Furhat, Galician language

1. Introduction

The advent of artificial intelligence has led to significant advancements in the field of social robots [1, 2, 3], which are characterized by their ability to interact and communicate with humans. The evolution of language technologies, particularly speech technologies, has been crucial in this progress, enabling robots to interact more naturally and effectively. These technologies include automatic speech recognition (ASR), text-to-speech synthesis (TTS), and dialogue systems, which empower robots to understand and respond to verbal commands, as well as engage in fluid conversations. Furthermore, many of these robots incorporate multimodal interaction capabilities, such as face recognition and face tracking systems, and enrich communication with the user through gestures and facial expressions.

Furhat [4] is a prime example of a cutting-edge social robot that has the capacity to modify its facial appearance and expressions through its innovative projected mask design. This technology enables it to interact with multiple people simultaneously in a multimodal approach, using verbal and non-verbal cues like speech, real-time face tracking, facial analysis, lip-synced facial animation, gestures, and eye and head movements, creating mixed-initiative conversations. In addition, Furhat integrates the most advanced ASR and TTS systems, supporting more than 40 languages and over 200 different voices.

However, to ensure a fully satisfying experience when interacting with social robots that feature human-like faces, attention must be paid to both the quality of synthesized speech and the synchronization of facial expressions. In particular, accurate lip synchronization is crucial for natural and fluid robot-human interaction, as it directly impacts the perceived intelligibility and expressiveness of the robot's speech. This synchronization must be carefully aligned with the prosody of the synthesized audio and the specific phonetic features of the target language.

Interacción '25: XXV International Conference on Human-Computer Interaction, September 03–05, 2025, Valladolid, Spain
*Corresponding author.
carlacastedo.pereira@usc.es (C. Castedo); mariadelcarmen.magarinos@usc.gal (C. Magariños); alejandro.catala@usc.es (A. Catala); alberto.bugarin@usc.es (A. Bugarín-Diz)

In this regard, Furhat's lip synchronization faces limitations due to its reliance on the Microsoft Universal Phone Set (UPS)[1], which is primarily designed for American English. This means Furhat cannot accurately reproduce lip movements for phones absent in the UPS, which is a significant issue when working with other languages. Moreover, integrating external TTS systems that lack the necessary information for lip synchronization (i.e. phone sequence and timestamps) presents an additional challenge. Furhat defaults to automatic lip-syncing based on phone audio recognition in such cases. This automatic process, while functional, is suboptimal, as the built-in phone recognizer is trained on English. Therefore, for languages other than English, the resulting lip movements are frequently inaccurate, further diminishing the naturalness and realism of Furhat's speech.

This work addresses the limitations of Furhat's default lip synchronization when integrating an external TTS system in the Galician language. We propose an alternative method that leverages the Montreal Forced Aligner (MFA) [5], a robust forced alignment tool based on a classical Hidden Markov Model-Gaussian mixture model (HMM-GMM) architecture, to generate accurate phone-level alignments for Galician speech. We evaluate this method using the Celtia voice of the Nós-TTS system [2] [6], a high-quality Galician synthetic voice developed within the Nós project [7, 8]. A key challenge overcome in this work is the creation of acoustic models and a pronunciation dictionary for Galician, as, to the best of our knowledge, no pre-existing resources were available. This study aims to bridge this gap, enabling the integration of MFA with Furhat and evaluating its performance in achieving accurate lip synchronization for Galician speech and improving user perception. Evaluation will be performed through objective measures and a perceptual preference test conducted by native speakers.

The remainder of this paper is structured as follows: Section 2 reviews background on Furhat's lip-syncing and forced alignment; Section 3 outlines the proposed system architecture; Section 4 details the employed methodology; Section 5 presents the evaluation results; and, finally, Section 6 discusses the conclusions and future research directions.

2. Background

2.1. Furhat's lip-syncing

Furhat's operation relies on Kotlin-based skills, which manage both speech output and facial expressions. Lip synchronization is achieved through the use of paired audio and alignment text files. Specifically, 16kHz 16-bit PCM WAV audio files are used, accompanied by corresponding Furhat-specific JSON-formatted text files (*.pho*), which provide precise word- and phone-level alignments. Additionally, each phone in a *.pho* file has an associated Boolean "prominent" field. This field is used to trigger a *MonitorSpeechProminent* event, which can be used for co-speech gestures like raising the eyebrows. However, the company does not provide specific code or criteria for its implementation.

In Furhat's speech generation process, the TTS system is expected to provide a *.pho* file for accurate lip synchronization. If this file is missing, Furhat employs an automatic lip-sync mechanism based on phonetic recognition of the input audio. While functional across languages, this automatic process is optimized for American English, as its underlying model is primarily trained on English data. Consequently, lip-syncing for other languages may be suboptimal, resulting in unnatural or inaccurate lip movements that do not correspond to the actual pronunciation. Therefore, to achieve accurate lip-syncing in other languages, a custom phonetic file that matches Furhat's requirements must be generated. For this purpose, a forced alignment tool may be used. The selected aligner should support phone-level alignment, and the resulting phonetic transcriptions must be mapped to the Microsoft UPS to ensure compatibility with Furhat.

Furhat Robotics also provides a web application [3] that uses the MFA for alignment of American English. Although this tool is not integrated for real-time operation with the robot, it has been used in this work to analyze the structure of the *.pho* files generated by Furhat.

[1] As per 2024-06-30, when this study ran.
[2] https://tts.nos.gal
[3] https://furhat.io/audio

2.2. Forced alignment

Forced alignment techniques can be categorized into three main approaches: HTK-based, Kaldi-based, and deep learning-based methods. The Hidden Markov Model Toolkit (HTK) [9] is an older yet widely used framework, powering aligners like Prosodylab-Aligner [10] and MAUS [11]. While these offer robust performance, they have limited flexibility and a steep learning curve. Kaldi-based [12] methods, exemplified by the MFA [5], overcome these limitations by offering a more user-friendly interface, active development, and more advanced acoustic modelling techniques. Deep learning-based methods, such as Wav2Vec 2.0 [13], employ self-supervised learning and Connectionist Temporal Classification (CTC) [14] to achieve high accuracy, often matching traditional methods performance [15, 16]. However, they typically require pre-existing alignments for training. Although NeMo Forced Aligner (NFA) [17] has recently emerged as a promising alternative, it does not provide phone-level alignment, making it unsuitable for Furhat's needs. Ultimately, MFA was chosen for this study due to its high accuracy [18], adaptability, noise resistance, ease of use, and comprehensive documentation.

MFA employs a traditional architecture that combines HMMs for sequence modelling with GMMs for acoustic modelling. The training process involves extracting Mel-Frequency Cepstral Coefficients (MFCC) from the audio and training phone models (monophones and triphones) using the Expectation-Maximization (EM) algorithm. To improve accuracy, speaker adaptation techniques, such as Linear Discriminant Analysis with a Maximum Likelihood Linear Transform (LDA+MLLT) and Speaker Adaptive Training (SAT) with Feature-space Maximum Likelihood Linear Regression (fMLLR), are applied. For alignment, the audio, the corresponding orthographic transcription, and a pronunciation dictionary are required. MFA supports the custom training of acoustic models and pronunciation dictionaries, facilitating the inclusion of new languages like Galician.

3. System overview

Figure 1 illustrates the system architecture for integrating forced alignment with the Furhat robot. The process begins with text input via a graphical interface or robot skill, which is then sent to the robot for speech synthesis. The robot then calls a proxy service that connects Furhat to external systems. This service makes two parallel requests: one to the TTS system for speech synthesis (returning a *.wav* file) and the other to the alignment service for the *.pho* file needed for lip synchronization.

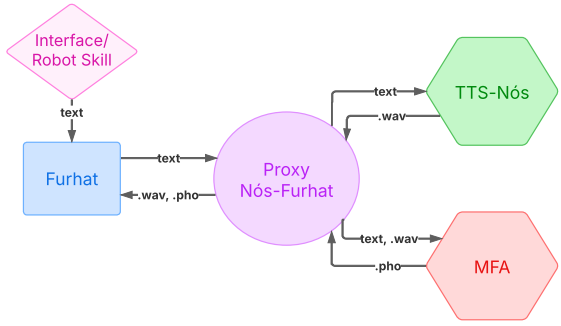


Figure 1: System architecture

In the default configuration, the proxy service responds only to the audio request. Therefore, the robot generates the *.pho* file internally. When using our proposed alignment method, the proxy service requests the *.pho* file from the alignment service, sending both the input text and the synthesized audio. The resulting *.pho* file is then sent to the robot, enabling enhanced lip synchronization. Our system requires a custom-built MFA module, including all necessary resources for forced alignment. Furthermore, a dedicated program, or "skill", must be running on the Furhat robot to manage these tasks.

4. Methodology

4.1. Pronunciation dictionary construction

The pronunciation dictionary was created using the established format for MFA dictionaries. Words and conjugated verb forms were drawn for the *Real Academia Galega* dictionary [19] and the *Instituto da Lingua Galega* pronunciation dictionary [20].

In addition, when validating a speech corpus with MFA, a list of out-of-vocabulary words (OOV) is obtained. These words belong to the corpus but are not included in the dictionary. In order to avoid them, the OOVs of all speech corpora that will be used later for the training of acoustic models [21, 22, 23, 24, 25] were included in the pronunciation dictionary. These words were normalized and phonetically transcribed using the text processing module of Cotovía [26], which serves as a grapheme-to-phoneme (G2P) system for the Galician language. Cotovía outputs use a special phone set designed for the tool. For simplicity, it was taken as the dictionary phone set. It must be noted that pronunciation probabilities were not considered, as the goal was the creation of a functional basic dictionary.

4.2. Acoustic model training

A custom acoustic model was trained to enable MFA to perform forced alignment in Galician. This process involved utilizing over 1,700 hours of speech data from various corpora. Due to the MFA's database size constraints, it was infeasible to train a model using the complete dataset. Therefore, several combinations using different corpora were used for training. The best results were obtained with the combination of Nós_Celtia-GL [22, 27], Common_Voice_GL_17.0 [21], Telexornais_LS (internal use) and OpenSLR77 [24], with a total of 216 hours of audio and more than 4,000 different speakers as shown in Table 1. The best results were determined based on the validation set described in Section 4.4 and the metrics detailed in Section 5.1.

Table 1
Hours, no. sentences, and no. speakers in speech corpora.

Corpus	Duration (h)	Sentences	Speakers
Nós_Celtia-GL	25	20,000	1
Nós_ParlaSpeech	1,197	667,308	216
Common_Voice_GL	66	44,427	2,264
Telexornais_LS	115	82,715	1,864
RG_Podcast	370	38,720	32
OpenSLR77	10	5,608	44
Total	1,783	858,778	4,421

4.3. Furhat's phonetic file construction

Because Furhat requires specific phonetic files for lip synchronization, the TextGrid format output by MFA, containing word- and phone-level alignments, must be converted to Furhat's *.pho* format. Furthermore, as previously mentioned, Furhat's phone articulation relies on the UPS. Although Furhat allows recording new gestures, it does not support modifications to its articulation phone set. Consequently, to generate a functional *.pho* file, a mapping is required between the Cotovía phone set (used for alignments) and the Furhat phone set. This mapping is performed via the International Phonetic Alphabet (IPA) [28], as detailed in Table 2. Due to incomplete equivalence between the phone sets, some phone approximations (highlighted in bold red) were made with the assistance of a phonetics expert. As for the prominent field, given the lack of information, we used the tonic syllable of each word as the criterion.

Table 2
Equivalence between Cotovía, IPA and UPS phone sets. Approximations for non-equivalent phones are highlighted in bold red.

Cotovia	a	E	e	i	j	O	o	u	w	p	b	B	t	d	D	k
IPA	a	ɛ	e	i	j	ɔ	o	u	w	p	b	β	t	d	ð	k
UPS	**AA**	EH	**EH**	I	J	AO	O	U	W	P	B	**B**	T	D	DH	K
Cotovia	g	G	f	T	s	S	C	m	n	N	J	l	Z	r	R	x
IPA	g	ɣ	f	θ	s	ʃ	tʃ	m	n	ɲ	ŋ	l	ʎ	ɾ	r	x
UPS	G	**G**	F	TH	S	SH	CH	M	N	N	NG	L	**JH**	**R**	**R**	**H**

4.4. Evaluation

In order to check whether the trained MFA model was suitable for the alignment of synthetic speech, a set of five sentences (included in Appendix A) was designed, considering particularly challenging articulations. These sentences, synthesized using the Celtia voice of the Nós-TTS system and manually aligned using Praat [29], will serve as a gold standard.

Using this gold standard as a reference, we performed an objective evaluation considering alignment errors between the timestamps predicted by the trained model and the manually annotated timestamps (at both word and phone levels). We calculated these errors' mean, standard deviation, and median to assess model performance and compare our results with those reported in the literature. The selected model is the one with the lowest statistic values, trained on the corpora specified in Section 4.2.

After validation, the alignment model was integrated into the proposed system. A subjective evaluation, by means of a perceptual preference test, was then conducted to compare the perception of the default lip-syncing and that obtained with our system. The test consisted of 10 pairs of stimuli, each corresponding to a different sentence. Within each pair, one stimulus featured Furhat's default lip-syncing, while the other showcased the synchronization generated by the proposed method. The stimuli in each pair were presented in random order to the participants. The selection of the 10 sentences, which comprised the 5 sentences from the objective evaluation, was guided by a phonetics expert. The sentences were designed to cover a range of expressiveness, including declarative, interrogative, and exclamatory forms. For each pair, participants were asked to indicate their preference by answering the question 'Which of the stimuli is more natural and synchronized with the audio?" The answer options were: "Stimulus A", "Stimulus B" and "I can't decide". Participants were allowed to ask for the stimulus of their choice to be replayed as many times as they wished. It is worth mentioning that participants experienced the stimuli [4] directly through the robotic head. The total task duration was approximately 20 minutes, and the test was conducted in a room free from noise interference.

5. Results

5.1. Automatic misalignment assessment

The evaluation of the Galician MFA-trained acoustic model was conducted by comparing its automatically generated alignments against the gold standard. Key metrics, including mean, median, and standard deviation, were calculated for both word- and phone-level alignment errors and are shown in Table 3.

At the word level, it can be observed that the results are consistent with those obtained in [15] in terms of mean, median and standard deviation. In that work, the authors used pre-trained acoustic models from the official MFA website. Therefore, the fact that our trained Galician model replicates their results suggests it performs comparably to other models provided by MFA in their acoustic model bank.

[4]Videorecordings are included in the following link to illustrate how the stimuli looked like. https://nextcloud.citius.usc.es/s/TNpGESXxnq3bmzB

Table 3
Mean, median and standard deviation of alignment errors

Level	Mean	Median	St. Dev.
Word	0.0673	0.0189	0.2113
Phone	0.0255	0.0092	0.1109

Concerning the errors at the phone level, our results are comparable to those reported by the authors of MFA [5] in terms of mean and median. This represents a significant achievement, considering the greater availability of high-quality, large-scale corpora for American English. The close results, despite limited access to corpora, suggest strong model performance.

It is worth remarking on the difference between the mean and median values, especially at the word level, as depicted in the boxplots in Figure 2a. As can be observed, while most data points cluster around the median for both word- and phone-level errors, a couple of outliers, deviating by more than one second, substantially increase the mean. These values, located around the same point, indicate that this large difference in phone alignment accumulates at the word level. Further investigation of the corresponding audio files revealed that these outliers stem from silence misalignments.

5.2. Perceptual preference test

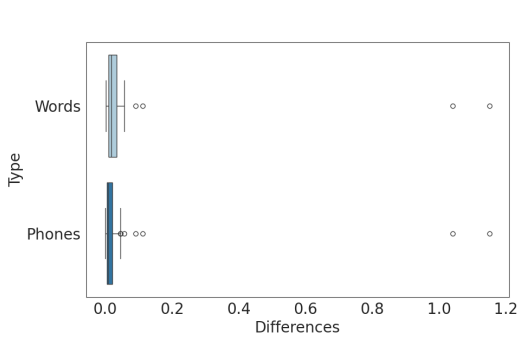
Five adult Galician native speakers, all women aged 20-40 with normal or corrected-to-normal vision, participated in the test. Four of them worked in language technologies, with backgrounds in fields such as linguistics, speech technologies, and computational linguistics. Three of these had prior experience with either Galician speech synthesis, the Furhat robot, or both. The fifth participant was completely unfamiliar with the technologies under consideration. Figure 2b shows the obtained preference scores for the default Furhat synchronization, the trained model and the situation of no preference for either, along with the corresponding 95% confidence intervals. Trained model synchronization was chosen in 88% of the cases, while in only 2% of them, the default Furhat synchronization was chosen. Moreover, confidence intervals do not overlap, revealing significant differences, even for such a low number of participants.

For 8 out of the 10 sentences, the preference was unanimous for the trained model. Moreover, in only one of the sentences, most participants did not choose the trained model. This sentence is also one of those considered in the previous objective evaluation, and it is the one that generated the outliers in the boxplots. When the participants were asked about this sentence, they stated that they were puzzled to see the robot moving its lips in a moment of silence. Thus, improved management of silences during model training could further increase the preference for the trained model. As reported by other authors in the literature [30], the lack of explicitly encoded silences in transcriptions poses a challenge for forced alignment, often leading to misalignment during long silent intervals. This issue could be mitigated by using the pause information provided by Cotovía to encode long silences in the transcription.

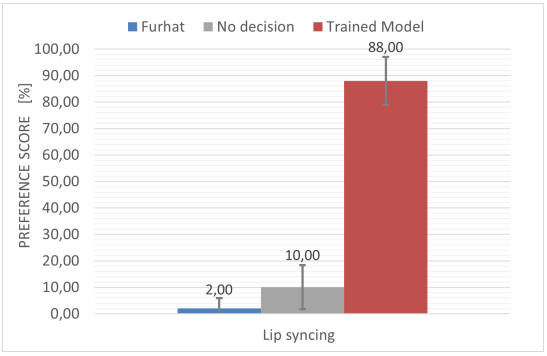
In addition, the phonetics expert was asked to provide her overall assessment of the synchronization. She stated that she perceived a significant improvement in the phone articulation when using the proposed method and even noted an increase in the robot's facial expressiveness. While further investigation is necessary, this enhancement in expressiveness may be related to the influence of the prominent field.

6. Conclusion

In this work we have explored an alternative form of lip-syncing for the social robot Furhat in its Galician speech by using Nós-TTS and forced alignment tools. Specifically, MFA was used to train an acoustic model for Galician completely from scratch, as well as build the pronunciation dictionary.



(a) Alignment errors at word and phone level



(b) Perceptual preference test results

Figure 2: Evaluation results

Automatic misalignment assessment demonstrates that our trained model is able to match, at the word level, the results obtained by pre-trained models. At the phone level, its performance is close to that reported for American English by the MFA developers.

The perceptual preference test shows the positive results achieved by the system with the integrated trained model. It is clear that the participants perceived better lip synchronization for the integrated system and that by adjusting the fields of the pronunciation file, it has been possible to gain expressiveness in facial expressions. These results are promising and indicate the feasibility of conducting further formal system validation with more participants. Additionally, it would be interesting to conduct a more detailed evaluation of the robot expressiveness associated to lip-syncing in particular social applications, where incorporating validated questionnaires or biometrics measures (e.g., eye-tracking, galvanic skin response) could offer more objective and in-depth insights on the interactions.

In light of the encouraging initial results reported in this paper, we intend to expand the scope of this study. Future work could evaluate the performance obtained by a probability-based pronunciation dictionary and explore alternative forced alignment systems, considering more recent deep learning-based models. Furthermore, we plan to enhance the implementation and deployment of the developed method to ensure lower latency, making it suitable for non-scripted scenarios outside the lab. Finally, given the importance of expressiveness, future studies could investigate alternative methods for defining the prominent field in order to achieve more natural and realistic gesticulation.

Acknowledgments

This work was funded by the Ministry for Digital Transformation and Civil Service and the Recovery, Transformation and Resilience Plan - Funded by EU – NextGenerationEU within the framework of the projects ”Desarrollo Modelos ALIA” and NEL-NÓS (ref. 2022/TL22/00215336). This research is also supported by projects PID2020-112623GB-I00, PID2021-123152OB-C21, and CNS2024-154915 funded by MCIN/AEI/10.13039/501100011033/ and by ERDF A way of making Europe. The support of the Galician Ministry for Education, Universities and Professional Training and the “ERDF A way of making Europe” is also acknowledged through grants “Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04” and “Reference Competitive Group accreditation 2022-2025 ED431C 2022/19”. Special thanks are extended to Albina Sarymsakova and Noelia García Díaz for their expert support on specific phonetic tasks.

Declaration on Generative AI

During the preparation of this manuscript, the authors used GPT-4, Gemini 2.0 Flash, and DeepL for text translation, grammar and spelling verification, and paraphrasing and rephrasing. After using

these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] T. Fong, I. Nourbakhsh, K. Dautenhahn, A survey of socially interactive robots, Robotics and Autonomous Systems 42 (2003) 143–166. doi:10.1016/S0921-8890(02)00372-X.

[2] A. Henschel, G. Laban, E. Cross, What Makes a Robot Social? A Review of Social Robots from Science Fiction to a Home or Hospital Near You, Current Robotics Reports 2 (2021). doi:10.1007/s43154-020-00035-0.

[3] H. Mahdi, S. A. Akgun, S. Saleh, K. Dautenhahn, A survey on the design and evolution of social robots — Past, present and future, Robotics and Autonomous Systems 156 (2022) 104193.

[4] S. Al Moubayed, J. Beskow, G. Skantze, The furhat social companion talking head, in: Proc. Interspeech 2013, 2013, pp. 747–749.

[5] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger, Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi, in: Proc. Interspeech 2017, 2017, pp. 498–502. doi:10.21437/Interspeech.2017-1386.

[6] C. Magariños, A. Öktem, A. Moscoso Sánchez, M. Vázquez Abuín, N. García Díaz, A. I. Vladu, E. Fernández Rei, M. Baqueiro Vidal, Nós-TTS: a Web User Interface for Galician Text-to-Speech, in: Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2, Association for Computational Linguistics, Santiago de Compostela, Galicia, Spain, 2024, p. 200–203.

[7] A. I. Vladu, I. de Dios-Flores, C. Magariños, J. E. Ortega, J. R. Pichel, M. Garcia, P. Gamallo, E. Fernández Rei, A. Bugarín, M. González González, S. Barro, X. L. Regueira, Proxecto Nós: Artificial intelligence at the service of the Galician language, in: SEPLN-PD 2022. Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations, A Coruña, Spain, 2022.

[8] I. de Dios-Flores, C. Magariños, A. I. Vladu, J. E. Ortega, J. R. Pichel, M. Garcia, P. Gamallo, E. Fernández Rei, A. Bugarín, M. González González, S. Barro, X. L. Regueira, The Nos' Project: Opening routes for the Galician language in the field of language technologies, in: Proceedings of the TDLE Workshop LREC2022, European Language Resources Association (ELRA), Marseille, 2022, pp. 52–61.

[9] S. Young, The HTK Hidden Markov Model Toolkit: Design and Philosophy, Entropic Cambridge Research Laboratory, Ltd 2 (1994) 2–44.

[10] K. Gorman, J. Howell, M. Wagner, Prosodylab-aligner: A tool for forced alignment of laboratory speech, in: Canadian Acoustics, volume 39, 2011, pp. 192–193.

[11] F. Schiel, Automatic Phonetic Transcription of Non-Prompted Speech, in: Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS), volume 1, 1999, pp. 607–610.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, The Kaldi Speech Recognition Toolkit, in: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, 2011. IEEE Catalog No.: CFP11SRW-USB.

[13] A. Baevski, H. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, in: Proceedings of the 34th Int. Conf. Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[14] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 23rd International Conference on Machine Learning, ICML '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 369–376. doi:10.1145/1143844.1143891.

[15] K. Biczysko, Automatic Annotation of Speech: Exploring Boundaries within Forced Alignment for Swedish and Norwegian, Master's thesis, Uppsala University, Dep. Linguistics and Philology, 2022.

[16] J. Zhu, C. Zhang, D. Jurgens, Phone-to-Audio Alignment without Text: A Semi-Supervised Approach, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8167–8171. doi:10.1109/ICASSP43922.2022.9746112.
[17] E. Rastorgueva, V. Lavrukhin, B. Ginsburg, NeMo Forced Aligner and its application to word alignment for subtitle generation, in: Proc. INTERSPEECH 2023, 2023, pp. 5257–5258.
[18] R. Rousso, E. Cohen, J. Keshet, E. Chodroff, Tradition or Innovation: A Comparison of Modern ASR Methods for Forced Alignment, in: Interspeech 2024, 2024, pp. 1525–1529. doi:10.21437/Interspeech.2024-429.
[19] Real Academia Galega, Dicionario da Real Academia Galega, 2024. URL: https://academia.gal/dicionario, [Accesed 2024-06-30].
[20] Instituto da Lingua Galega, Universidade de Santiago de Compostela, Dicionario de Pronuncia da Lingua Galega, 2024. URL: http://ilg.usc.es/pronuncia/, [Accesed 2024-06-30].
[21] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, G. Weber, Common Voice: A Massively-Multilingual Speech Corpus, in: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2020, pp. 4211–4215.
[22] M. Vázquez Abuín, N. García Díaz, A. I. Vladu, C. Magariños, A. Vidal Miguéns, E. Fernández Rei, Nos_Celtia-GL: Galician TTS corpus, 2023. doi:10.5281/zenodo.7716958, dataset.
[23] C. Magariños, A. V. Miguéns, A. I. Vladu, N. G. Díaz, M. V. Abuín, A. V. Couso, D. Bardanca, E. F. Rei, Nos_ParlaSpeech-GL: Galician ASR corpus, 2023. doi:10.5281/zenodo.7913218, dataset.
[24] O. Kjartansson, A. Gutkin, A. Butryna, I. Demirsahin, C. Rivera, Open-Source High Quality Speech Datasets for Basque, Catalan and Galician, in: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), European Language Resources association (ELRA), Marseille, France, 2020, pp. 21–27.
[25] C. Canosa, J. J. Francisco, A. Moscoso, J. González Corbelle, N. García, C. Magariños, A. I. Vladu, D. Fernández López, E. Fernández Rei, F. Dubert-García, X. L. Regueira, Nos_RG-Podcast-GL, 2025. URL: https://huggingface.co/datasets/proxectonos/Nos_RG-Podcast-GL, [Accesed 2024-06-30].
[26] E. Rodríguez Banga, C. García-Mateo, F. Méndez-Pazó, M. González-González, C. Magariños, Cotovía: an open source TTS for Galician and Spanish, in: VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, IberSPEECH, 2012, pp. 308–315.
[27] N. García Díaz, M. Vázquez Abuín, C. Magariños, A. I. Vladu, A. Moscoso Sánchez, E. Fernández Rei, Nos_Celtia-GL: an Open High-Quality Speech Synthesis Resource for Galician, in: IberSPEECH 2024, 2024, pp. 91–95. doi:10.21437/IberSPEECH.2024-19.
[28] International Phonetic Association, International Phonetic Alphabet Chart, 2020. URL: https://www.internationalphoneticassociation.org/content/ipa-chart, accessed 2024-06-24.
[29] P. Boersma, D. Weenink, Praat: doing phonetics by computer, https://www.fon.hum.uva.nl/praat/, 1992. [Accesed 2024-06-30, version 6.4.13].
[30] J. Zhu, C. Zhang, D. Jurgens, Phone-to-audio alignment without text: A semi-supervised approach, 2022, pp. 8167–8171. doi:10.1109/ICASSP43922.2022.9746112.

A. Set of Sentences Used for Evaluation

Objective Evaluation

1. Partiu a leña do souto.
2. Vén aquí para velo ben.
3. Que dis? É difícil, non?
4. A cadela desenterrou os ósos.
5. Cala, ho! Non deixas escoitar.

Additional Sentences for Perceptive Test

6. Unha curuxa! Berrou o vello.
7. Esa muller e o seu home son meus veciños.
8. O proxecto foi nado na internet galaica.
9. Choveu moito onte en Santiago.
10. Centos de veces merquei ese xornal.