Forced Phonetic Alignment in Brazilian Portuguese Using Time-Delay Neural Networks

Cassio Batista(✉) and Nelson Neto

Computer Science Graduate Program, Federal University of Pará, Belém, Brazil
{cassiotb,nelsonneto}@ufpa.br

Abstract. Forced phonetic alignment (FPA) is the task of assessing the time boundaries of phonetic units, i.e., calculating when in the speech utterance a certain phoneme starts and ends. This paper describes experiments on FPA for Brazilian Portuguese using Kaldi toolkit. Based on time-delay neural networks (TDNN), several acoustic models were trained on the top of the combination between hidden Markov models (HMM) and Gaussian mixture models (GMM). The nature of the input features and the topology of the HMMs have been varied in order to analyze each one's influence. Results with respect to the phone boundary metric over a dataset of 385 hand-aligned utterances show that the network is mostly invariant to the input features, while regular HMM topologies do perform better in comparison to a modified version used in chain models. Conversely, the neural network still does not outperform GMM models for phonetic alignment.

Keywords: Forced phonetic alignment · Speech segmentation · Acoustic modeling · Kaldi · Brazilian Portuguese

1 Introduction

Forced phonetic alignment (FPA) is the task of aligning a speech recording with its phonetic transcription, which is useful across a myriad of linguistic tasks. However, annotating phonetic boundaries of several hours of speech by hand is very time-consuming, even for experienced phoneticians. As several approaches have been applied to automate this process, some of them brought from the automatic speech recognition (ASR) domain, the combination of hidden Markov models (HMM) and Gaussian mixture models (GMM) has been for long the most widely explored for FPA.

Before Kaldi's success on ASR [10], mainly due to its efficient implementation of deep neural networks (DNN) for HMM-DNN hybrid acoustic modeling, EasyAlign [4] and UFPAlign [15] were the only ASR-based forced aligners with support to Brazilian Portuguese (BP), based on HTK toolkit [16]. Nowadays, Montreal Forced Aligner (MFA) [8] and an updated version of UFPAlign [1,3] both provide Kaldi-compliant acoustic models for BP.

V. Pinheiro et al. (Eds.): PROPOR 2022, LNAI 13208, pp. 323–332, 2022.
https://doi.org/10.1007/978-3-030-98305-5_30

Hence, this work provides an additional study on phonetic alignment using Kaldi tools, bearing in mind the idea of seeking improvements using the default neural network architecture that achieves state of the art for ASR in Kaldi's `nnet3` framework—factorized time-delay neural networks (TDNN-F) [9]. In total, 24 acoustic models were trained by varying i) the GMM model the network was trained upon (monophones, triphones and speaker-adapted triphones, namely `mono`, `tri-deltas` and `tri-sat`); ii) the HMM topology used as reference (single state vs. three-state, namely `chain` or `no-chain`); iii) and finally the input features: MFCCs or LDA, with or without stacked i-vectors.

To overcome the time it would take to train all such models, only LaPS-Benchmark dataset was used, which sums up to a total of only 54 min of recorded speech, 700 utterances divided among 10 female and 25 male native BP speakers. Evaluation procedures, on the other hand, were performed over 385 manually aligned audio files, 193 spoken by a male speaker and 192 spoken by a female speaker, which sums up to about 15 min of speech. The similarity measure is given by the absolute difference between the forced alignments with respect to manual ones, which is called phonetic boundary [8]. All scripts and resources have been release under MIT open license on GitHub[1].

2 Model Training and Evaluation Tests

The deep-learning-based training approach in Kaldi actually uses the GMM training as a pre-processing stage. For details on the GMM training pipeline, the reader is referred to [3]. The DNN is trained on the top of the last GMM model of the pipeline, which usually comprises a speaker-adapted triphone training (SAT). However, we experimented with monophones and triphones without SAT as well.

Figure 1 details how the DNN model is obtained as a final-stage acoustic model (AM) by using the neural network to model the state likelihood distributions as well as to input those likelihoods into the decision tree leaf nodes [5]. The implementation in Kaldi uses a sub-sampling technique (with a default factor of 3) that avoids the whole computation of a feed-forward's hidden activations at all time steps and therefore allows a faster training of TDNNs.

Models were trained following Kaldi's LibriSpeech recipe. The hardware setup consists of an Intel® Core™ i7-10700 octa-core processor, 32 GB of RAM and an 10 GB NVIDIA GeForce RTX 3080 GPU running CUDA 11.3. Although in Kaldi the number of epochs actually differs from the number of iterations in which the algorithm "sees" each data point, the former was set to 10 for all simulations. The number of layers was reduced to seven, as opposed to 16 in the default recipe for speeding up purposes. Layers were kept with dimension 1,536 each. Time strides were also left as is, with three past frames and nine future frames (i.e., left and right context w.r.t. the reference frame, respectively.) As the amount of training data was made to be limited, all five HMM-GMM models and 24 HMM-DNN could be trained in less than 12 h.

[1] https://github.com/falabrasil.

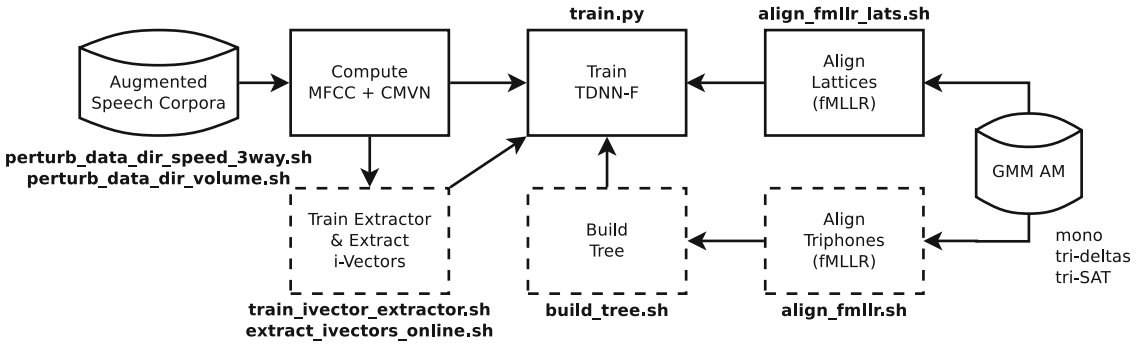


Fig. 1. Stages for training a TDNN-F in Kaldi. On the left side, high-resolution, cepstral-normalized MFCCs (40 features instead of 13) are extracted from an augmented corpora after applying speed and volume perturbation [7], as are the speaker-related 100-dimensional i-vectors [2,14]; to be used as input to the neural network. On the right side, labels are provided by a GMM acoustic model. Dashed blocks may or may not be used. For instance, the tree must only be rebuilt when using `chain` models, in which the HMM topology is modified, whereas i-vectors may be excluded from the network's input as well.

The evaluation procedure takes place by comparing pairs of annotated files: the alignments that we consider as gold standard (hand-aligned reference), and the ones automatically annotated by inference (forced-aligned hypothesis). The phone boundary metric considers the absolute difference between the ending time of both phoneme occurrences [8], as a phone's beginning time is considered the same as its predecessor's ending time. The calculation is performed for each acoustic model, and involves all utterances from the evaluation dataset composed by one male and one female speaker.

2.1 Evaluation Speech Corpus

The automatic alignment was estimated on the basis of the manual segmentation. The original dataset used for assessing the accuracy of the phonetic aligner is composed of 200 and 199 utterances spoken by a male and a female speaker, in a total of 15 min and 32 s of hand-aligned audio, as shown in Table 1. Praat's TextGrid files, whose phonetic timestamps were manually adjusted by a phonetician, are available alongside audio and text transcriptions.

The authors acknowledge that this corpus is rather limited in both its size and speaker variability: 15 min of recorded speech from two speakers is too small indeed. However, we emphasize the difficulty to get access to this kind of labeled data, which very often requires an expert phonetician to spend hours aligning audios by hand.

This dataset was aligned with a set of phonemes inspired by the SAMPA alphabet, which in theory is the same set used by the FalaBrasil's G2P software that creates the lexicon during acoustic model training. Nevertheless, there are some problems of phonetic mismatches, and some cross-word phonemes between

Table 1. Speech corpus used to evaluate the automatic phonetic aligners. Actual duration and number of files after discard are shown between parentheses, as well as the number of unique words.

Dataset	Duration	# Files	# Words	# Tokens
Male	7 m:58 s (7 m:40 s)	200 (193)	1,260 (665)	5,275
Female	7 m:34 s (7 m:18 s)	199 (192)	1,258 (664)	5,262
Total	15 m:32 s (14 m:58 s)	399 (385)	2,518 (686)	10,537

words, which makes the mapping between both phoneme sets challenging, given that FalaBrasil's G2P only handles internal-word conversion [13].

The example in Table 2 shows the phonetic transcription for a sentence given by the original dataset (top) and the acoustic model (bottom) which then suppress vowel sounds altogether due to cross-word rules (usually elision and apocope) when they occur at the end of the current word and at the beginning at the next. Such mismatches occur because the dataset was aligned by a phonetician considering acoustic information (i.e., listening), which cannot be done by the G2P tool that creates the acoustic model's lexicon, since it is provided only with textual information. Situations like these of phonetic information loss led to the removal of such audio files from the dataset before evaluation.

Table 2. Cross-word mismatches between transcriptions manually aligned by a phonetician (top) vs. generated by FalaBrasil's G2P software (bottom). Word boundary losses are represented by the empty set symbol (∅).

(a) "***ás*** ***nov*i** *meia*, ***pair*um** ***ar*** *no rio*" → "***ás*** ***nov*e** **e** *meia*, ***pair*a** **um** ***ar*** *no rio*"

6	∅	Z	n	O	v	i	∅	...	p	a	j	4	∅	u~	m	a	h/	...
a	j	s	n	O	v	i	i	...	p	a	j	r	a	u~	∅	a	X	...

In the end, fourteen files were excluded from the dataset, so about 34 s of audio was discarded, and 193 and 192 utterances remained in the male and female datasets, respectively. The filtering also ignored intra- and inter-word pauses and silences, resulting in 2,518 words (686 unique, since the utterances' transcriptions are identical for both speakers, i.e., they speak the very same sentences) and 10,537 phonetic segments (tokens) (c.f. Table 1).

2.2 Simulation Overview

Figure 2 shows a diagram of the experiments where the input audio files (`.wav`) with their respective textual transcriptions (`.txt`) are passed to Kaldi aligner. These are the files whose manual annotation is available. The output is a TextGrid file (`.tg`) for each audio given as input, which then serve as the inference inputs to the phone boundary calculation. The reference ground-truth annotations, on the

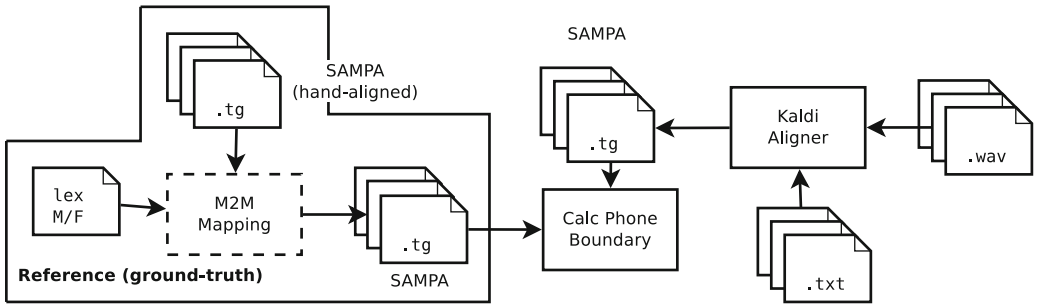


Fig. 2. Evaluation takes place by comparing the output of Kaldi to a hand-aligned ground-truth. The M2M mapping is applied to make different phone sets match the SAMPA version used by FalaBrasil's G2P, which is provided by the lexicon generated over transcriptions of the corpus (`lex M/F`).

other hand, are provided by the 385 TextGrid files that contain the hand-aligned phonemes corresponding to the transcriptions in the evaluation dataset.

However, for computing phone boundaries, there must exist a one-to-one mapping between the reference and the inference phones, which was not possible at first due to the nature of the phonetic alphabets: in our experiments, we used the SAMPA-inspired lexicon generated by FalaBrasil's G2P tool, while the hand-aligned utterances (referred here as "original") are also available in a SAMPA-inspired phonetic alphabet, but not exactly the same as FalaBrasil's.

Apart from the fact that cross-word rules can insert or delete phones, some phonemes do not have an equivalent, such as `/tS/` and `/dZ/`. Besides, there are also usual swaps between phonetically similar sounds: `/h//`, `/h\/`, `/h/` and `/4/`, for instance, might be almost deliberately mapped to either `/r/`, `/R/` or `/X/`.

Thus, since the situation seemed to require a smarter approach than a simple one-to-one tabular, static mapping, it was necessary to employ a many-to-many (M2M) mapping procedure (c.f. dashed blocks on Fig. 2) based on statistical frequency of occurrence, e.g., how many times phones `/t/` and `/S/` from the original evaluation dataset were mapped to a single phone `/tS/` in the `lex M/F` file representing FalaBrasil's G2P SAMPA-inspired alphabet.

By taking another look at Table 2, one might have also reasoned that the mapping between the two sets of phonemes is not always one-to-one. The usual situation is where a pair of phonemes from the dataset (original) is merged into a single one for the AM (FalaBrasil G2P), such as `/i~/ /n/` → `/i~/` and `/t/ /S/` → `/tS/`. However, a single phoneme can also be less frequently split into two or more, such as `/u/ /S/` → `/u/ /j/ /s/`.

To deal with these irregularities, we used the many-to-many alignment (m2m-aligner) software [6] in the core of a pipeline that converts the original TextGrid from the evaluation dataset to a TextGrid that is compatible with the FalaBrasil's lexicon used to train the acoustic models. The m2m-aligner works in an unsupervised fashion, using an edit-distance-based algorithm to align two strings in order for them to share the same length [6].

3 Results and Discussion

Results will be reported in terms of tolerance thresholds of 10, 25 and 50 ms that show how many phonetic tokens were more precisely aligned with respect to the manual alignments in the context of the phone boundary metric. Numeric values are shown in Figs. 3 and 4 for the GMM- and TDNN-F-based models, respectively. All bar charts have been trimmed at 30% percentage for the sake of a better visualization. Both graphs report average measures of five independent training repetitions of each acoustic model, hence vertical caps/whiskers represent the standard deviation. Blue- and red-shaded bars represent the male and the female speakers from the evaluation dataset, respectively.

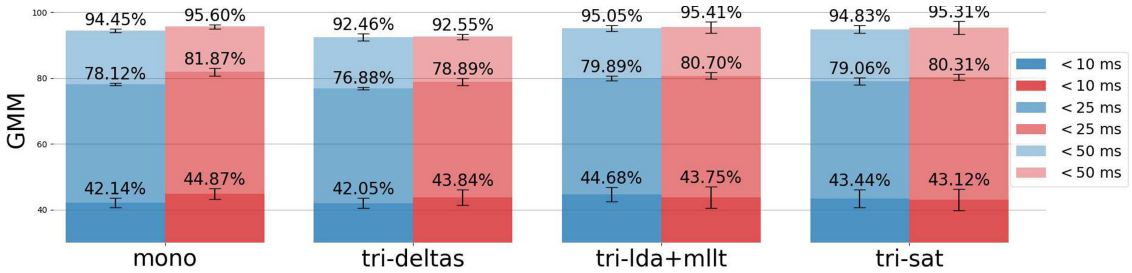
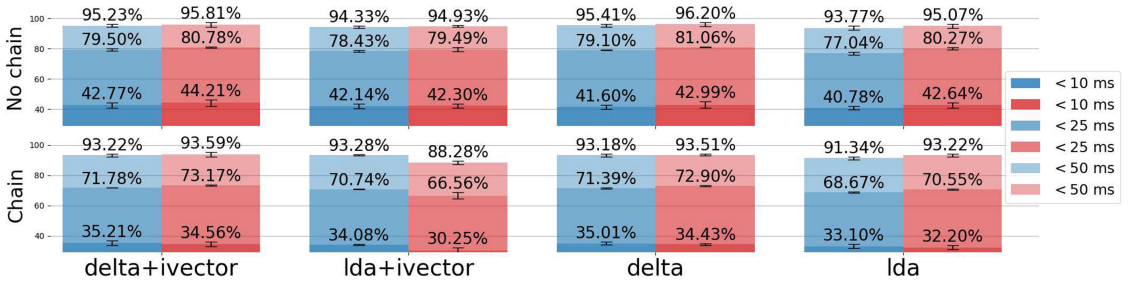


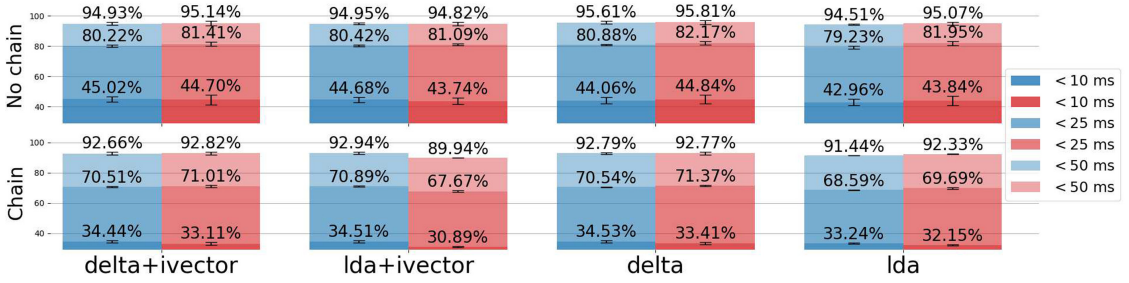
Fig. 3. Cumulative percentage below a tolerance threshold, in milliseconds, of the differences between forced aligned audio and ground-truth (hand aligned) phonemes, also known as phone boundary. Results for monophones and triphones trained with low resolution, MFCCs (deltas), spliced-MFCCs (LDA+MLLT) and speaker adaptation (SAT) within the HMM-GMM framework. (Color figure online)

The performance of the GMM models is shown in Fig. 3. As it can be seen, all four models' behaviors were virtually the very same: there is nearly a 1% difference in tokens correctly aligned within the 10 ms threshold, with an average of ∼43% across all models. With respect to the gender of the speaker, there have not been much of a difference either. One could say that tri-SAT and tri-LDA+MLLT contain all the high numbers for phone boundary, but we would rather state their performance is comparable. At the tolerance of 50 ms, more than 90% of the phonemes are correctly aligned with the speech.

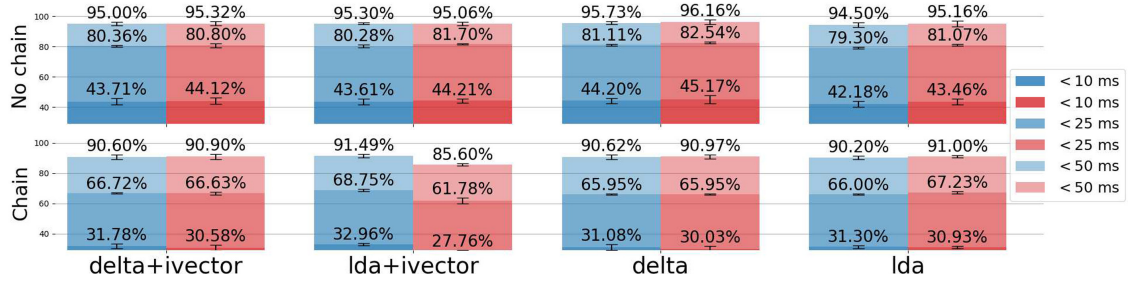
Figure 4 shows all results for the TDNN-F models. Again, at the 10 ms threshold, one cannot observe a significant difference among distinct models and input features. Stacking i-vectors on the top or splicing MFCCs with linear discriminant analysis procedures do not seem to improve performance at all. Topology-wise, on the other hand, we can say that the design of the HMMs under the chain modeling framework is not well suited for phonetic alignment. As with regular three-state left-to-right HMM topology (i.e., no chain) [11] the range of the cumulative percentages stayed within the 41%–45%, similar to the obtained with GMM models, it dropped to around the 30% in the chain models. For higher tolerance values, models outside the chain framework achieved the average of 80% and 94% of tokens correctly aligned, while chain models achieved only ∼65% and ∼90%, respectively.



(a) TDNN-F over monophones



(b) TDNN-F over tri-deltas



(c) TDNN-F over tri-SAT

Fig. 4. Results for the TDNN-F trained on the top (hybrid HMM-DNN framework) of the monophones, tri-deltas and tri-SAT GMM-based AMs using either MFCCs (delta) or spliced-MFCCs (LDA) as input, with or without i-vectors. (Color figure online)

Overall, on the chain-free framework, delta features (MFCCs) achieved the highest percentages for both speakers, sometimes combined with i-vectors as well. Regarding the GMM-based model upon which the TDNN-F is trained, tri-SAT models provide the best results for the female speaker, while tri-deltas was best for the male speaker. But once again, as maximum difference observed is 2%, the gains are so marginal that one should avoid using the word "outperform".

An interesting point is that one can achieve good results even with as little as one hour of recorded speech from LaPSBenchmark to train a model for phonetic alignment: past experiments on a dataset of ∼170 h of recorded speech also show results floating around the 45% percentage at 10 ms [1]. What we cannot tell, however, is how would the network behave if the training was scaled up to an 1,000 h dataset. MFA, for instance, uses the GlobalPhone dataset [12], which we consider already small although it contains 22 h for Brazilian Portuguese. As a matter of fact, MFA also uses GMM models, which may emphasize that

navigating through all the burden to train any DNN model (which requires at least one GPU card) may not be the more appropriate move if the final task's goal is to align phonemes rather than to recognize speech.

Furthermore, one downside is that most pre-trained models available for Kaldi are based on the chain framework for speech recognition, and apart from MFA releases there is a shortage on releases of GMM-based models as they are no longer useful for ASR. One thing to keep in mind would be to train models that perform relatively well for both tasks altogether—ASR and FPA.

4 Conclusion

This paper presented experiments on forced phonetic alignment (FPA) in Brazilian Portuguese (BP). In total, 24 acoustic models trained with Kaldi over TDNN-F networks (which represent the default architecture in Kaldi for state of the art speech recognition in the so-called conventional or hybrid approach) were evaluated. After tests considering the phone boundary metric, we found that the default chain models performed worse, probably because of their simplified HMM topology in the decision tree [11], while chain-free models and GMM are comparable. Scripts to train the models and other resources have been released under MIT open license on GitHub[2].

We understand the limitations of using only LapsBenchmark, which contains only 54 min of recorded speech, to train complex models based on deep neural networks such as TDNN-F. Therefore, as future work, we expect to extend the simulations with non-chain models over all public audio datasets for BP that have been recently released in order to verify whether a more robust model can serve both speech recognition and forced alignment tasks. We also plan on exploring the use of frame-subsampling on chain models, since the TDNN-F is configured to only "see" a third of the frames. We suspect that if this feature is disabled, the alignments could be more accurate. Moreover, other architectures like LSTMs or CNNs in combination with time-delay networks should have its use evaluated.

Finally, the employment of transfer learning techniques could be investigated to take advantage of the abundance of audio data for other languages like English. That way, an acoustic model trained over LibriSpeech dataset, for example, could be trained to serve as a starting point, and GMM-based models would be trained from scratch over the male/female evaluation dataset to play the role of the new tri-SAT reference alignments. Once more, the impediment, however, is that most of the pre-trained TDNN-F-based models available on the Internet are chain models, so a chain-free would have to be trained from scratch. Additionally, other newer toolkits could be explored alongside Kaldi if configured in a phoneme-based setup, like the new K2-Lhotse-Icefall trilogy, ESPnet or SpeechBrain.

[2] https://github.com/falabrasil.

Acknowledgment. We gratefully acknowledge NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. The authors also would like to thank CAPES for providing scholarships and FAPESPA (grant 001/2020, process 2019/583359) for the financial support.

References

1. Batista, C., Neto, N.: Experiments on kaldi-based forced phonetic alignment for brazilian portuguese. In: Britto, A., Valdivia Delgado, K. (eds.) Intelligent Systems, pp. 465–479. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-91699-2_32
2. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. **19**(4), 788–798 (2011). https://doi.org/10.1109/TASL.2010.2064307
3. Dias, A.L., Batista, C., Santana, D., Neto, N.: Towards a free, forced phonetic aligner for brazilian portuguese using kaldi tools. In: Cerri, R., Prati, R.C. (eds.) Intelligent Systems, pp. 621–635. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-61377-8_44
4. Goldman, J.P.: Easyalign: an automatic phonetic alignment tool under praat. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 3233–3236 (2011)
5. Guiroy, S., de Cordoba, R., Villegas, A.: Application of the kaldi toolkit for continuous speech recognition using hidden-markov models and deep neural networks. In: IberSPEECH'2016 On-line proceedings, pp. 187–196. IberSPEECH 2016, Lisboa, Portugal (2016)
6. Jiampojamarn, S., Kondrak, G., Sherif, T.: Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pp. 372–379. Association for Computational Linguistics, Rochester, New York (2007)
7. Ko, T., Peddinti, V., Povey, D., Khudanpur, S.: Audio augmentation for speech recognition. In: Proceedings of Interspeech (2015)
8. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M.: Montreal forced aligner: trainable text-speech alignment using kaldi. In: Proceedings of Interspeech, pp. 498–502 (2017). https://doi.org/10.21437/Interspeech.2017-1386
9. Peddinti, V., Povey, D., Khudanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts. In: Proceedings of Interspeech, pp. 3214–3218 (2015)
10. Povey, D., et al.: The kaldi speech recognition toolkit. In: In IEEE 2011 workshop (2011)
11. Povey, D., et al.: Purely sequence-trained neural networks for ASR based on lattice-free mmi. In: Interspeech 2016, pp. 2751–2755 (2016). https://doi.org/10.21437/Interspeech.2016-595
12. Schultz, T., Vu, N.T., Schlippe, T.: Globalphone: a multilingual text speech database in 20 languages. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8126–8130 (2013). https://doi.org/10.1109/ICASSP.2013.6639248
13. Siravenha, A., Neto, N., Macedo, V., Klautau, A.: Uso de regras fonológicas com determinação de vogal tônica para conversão grafema-fone em português brasileiro (2008)

14. Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S.: Deep neural network embeddings for text-independent speaker verification. In: Proceedings Interspeech 2017, pp. 999–1003 (2017). https://doi.org/10.21437/Interspeech.2017-620
15. Souza, G., Neto, N.: An automatic phonetic aligner for brazilian portuguese with a praat interface. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (eds.) Computational Processing of the Portuguese Language, pp. 374–384. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41552-9_38
16. Young, S., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book. Cambridge University Engineering Department, version 3.4, Cambridge, UK (2006)