IberSPEECH 2022
14-16 November 2022, Granada, Spain

GTTS Systems for the Albayzin 2022 Text and Speech Alignment Challenge

Germán Bordel, Luis J. Rodriguez-Fuentes, Mikel Peñagarikano, Amparo Varona

Department of Electricity and Electronics
Faculty of Science and Technology, UPV/EHU
Barrio Sarriena, 48940 Leioa, Spain.
{german.bordel, luisjavier.rodriguez, mikel.penagarikano, amparo.varona}@ehu.eus

Abstract

This paper describes the most relevant features of the alignment approach used by our research group (GTTS) for the Albayzin 2022 Text and Speech Alignment Challenge: Alignment of re-spoken subtitles (TaSAC-ST). It also presents and analyzes the results obtained by our primary and contrastive systems, focusing on the variability observed in the RTVE broadcasts used for this evaluation. The task is to provide some hypothesized start and end times for each subtitle to be aligned. To that end, our systems decode the audio at the phonetic level using acoustic models trained on external (non-RTVE) data, then align the recognized sequence of phones with the phonetic transcription of the corresponding text and transfer the timestamps of the recognized phones to the aligned text. The alignment error for each subtitle is computed as the sum of the absolute values of the start and end alignment errors (with regard to a manually supervised ground truth). The median of the alignment errors (MAE) for each broadcast is reported to compare system performance. Our primary system yielded MAEs between 0.20 and 0.36 seconds on the development set, and between 0.22 and 1.30 seconds on the test set, with average MAEs of 0.295 and 0.395, respectively.

Index Terms: Text and Speech Alignment, Automatic Speech Recognition, Cross-Domain Acoustic Models

1. Introduction

The Albayzin 2022 Text and Speech Alignment Challenge, which is part of the Albayzin 2022 Evaluation Campaign [1], proposed two different alignment tasks: (1) the alignment of re-spoken subtitles of RTVE broadcasts, which is concisely called TaSAC-ST [2]; and (2) the alignment of Basque Parliament plenary session minutes, called TaSAC-BP [3]. The two tasks were closely related but differed in a number of aspects: channel and background/environment conditions, number and type of speakers, required detail of the alignments and performance metrics. Our research group (Grupo de Trabajo en Tecnologías Software, GTTS), as the organizer of TASAC-BP, developed an alignment system specifically optimized for that evaluation, with the aim to provide a baseline to potential participants. It was only a few days before the submission deadline that we decided to adapt the baseline system developed for TASAC-BP to the TASAC-ST evaluation, with just the required adjustements to meet the evaluation conditions (output file format, etc.).

An important issue with our submission to TASAC-ST was that our acoustic models had been trained on cross-domain (non-RTVE) materials, differing from the target audios in different regards: channel, background/environment conditions, speakers and even the spoken languages. This may seriously hinder the ability of our systems to decode the TV broadcasts on which this evaluation remains. On the positive side, our proposed method may still have margin for improvement if the provided RTVE materials were used for training. Our original submission consisted of two systems (primary and contrastive-1), which applied the same approach but different acoustic models (trained on two independent sets of non-RTVE data). Two late (post-key) systems (contrastive-2 and contrastive-3) have been also submitted to the evaluation, showing improved performance thanks to a kernel modification in our dynamic programming algorithm which provides more compact alignment hypotheses.

The paper is organized as follows. Section 2 summarizes the main features of the Albayzin 2022 Text and Speech Alignment Challenge on re-spoken subtitles. Section 3 describes our alignment approach, with details about the phone decoder, the training datasets and the modified kernel used in our late submission. Section 4 presents and briefly discusses the obtained results, and Section 5 summarizes our contribution and points out avenues for future work.

2. The text and speech alignment task

To carry out the alignment task, a set of audio files in AAC format along with the corresponding set of text (UTF-8) files in STM format are provided. The STM format specifies one line per subtitle with 7 items per line: file name (without extension), channel identifier, speaker identifier, start time (in seconds from the beginning of the audio file), end time, label and subtitle text.

For this evaluation, only the audio file name, the start and end timestamps, and the subtitle text are relevant. The timestamps provided by the organizers form an increasing sequence of real numbers but are wrong. The task consists of processing the audio file, synchronize its contents with the subtitles and produce a new STM file which includes the start and end timestamps obtained by the alignment system for each subtitle. In all other respects (for example, the subtitle text), the output STM file must be identical to the original STM file.

The audio files and the subtitles used in this evaluation were extracted from diverse kinds of TV programs. The subtitles, which were generated by re-speaking when those programs were broadcast for the first time, do not always match the spoken audio, sometimes being reduced or even paraphrased. This adds difficulty to the alignment process.

As a part of the RTVE2022 database created to support the Albayzin 2022 evaluation challenges, two sets of audios and STM files were provided to tune and evaluate the alignment systems: a development set including 4 audios recorded from two RTVE programs, lasting 2 hours and 10 minutes; and a test set including 22 audios recorded from three RTVE programs (the two already used in the development set plus a new one), lasting 12 hours and 10 minutes.

10.21437/IberSPEECH.2022-58

The organizers also provided the participants with training data (speech and text) to estimate acoustic and language models. These data (the RTVE2018 and RTVE2020 databases) come from two previous Albayzin evaluation challenges, and include 164 hours of audio with human supervised transcriptions and 460 hours of audio with just the subtitles (not supervised by humans), along with texts extracted from subtitles of the 24H news channel. Participants were allowed to use other data for training their systems, as long as enough and suitable information (size, origin or name of the database, etc.) was provided in the description paper.

System performance was measured in terms of the absolute value of the difference between the timestamps provided by the system and the reference (manually generated) timestamps. For each subtitle i, the start and end time errors were computed as follows:

$$e_i = abs(t_i^{(hyp)} - t_i^{(ref)}) \quad (1)$$

The alignment error for each subtitle was defined as the sum of the start and end time errors. Then, for each audio file to be aligned, the median of the alignment errors (MAE), also known as Program Time-Error metric (PTEM), was used as performance score. Finally, the average of the MAE's (also known as Average PTEM) obtained on the audio files included in the test set was used as system score. For more details, see [2].

3. The GTTS alignment approach

In our approach, we do not use any language nor phonological models at all. We rely on acoustic models to perform an unrestricted phone decoding of the audio signal. Given an audio file X and the corresponding STM file with the subtitles S, a phone decoder is applied to X which produces a recognized sequence of phone-like units p_X (with timing information attached), and a grapheme-to-phoneme (G2P) converter is applied to S which produces a reference sequence of phone-like units p_S (with word and subtitle information attached). Then, the two sequences of phone-like units are aligned under the criterion of maximizing the number of matches in the alignment path, following the same alignment method that is being succesfully applied by our group for the synchronization of BP subtitles [4] [5] [6]. As a result of this process, the timing information is transferred from p_X to p_S. Finally, a new STM file is created, identical to the source STM except for the timestamps, which are obtained from the alignment: for each subtitle, the start time of the first word and the end time of the last word are used as start and end timestamps (see Figure 1).

3.1. The phone decoder

For this evaluation, we have applied the bilingual Basque-Spanish phone decoder that we currently use to process the audios and minutes of the Basque Parliament (BP) plenary sessions and produce synchronized subtitles. A reduced inventory of 23 acoustic units is used which suitably covers the most common sounds in both languages, along with an additional unit which accounts for silences and other non-linguistic events (see [7] for details).

In fact, we have applied two different phone decoders. The first one (used for our primary and contrastive-2 systems) was trained on external (non-BP) data in Basque and Spanish: CommonVoice (cv-corpus-5.1-2020-06-22, Basque and Spanish subsets) [8], OpenSLR (SLR76) [9], Aditu [10] and Albayzin [11]. Overall, these datasets amount to 332 hours of speech, with a high imbalance of Spanish over Basque (with a 3:1 ratio).

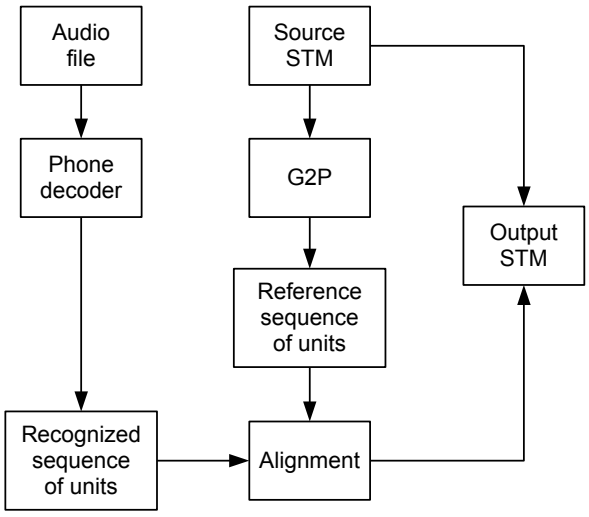

Figure 1: *After aligning the phone sequences p_X (recognized) and p_S (reference), the timing information attached to p_X is transferred to the words/subtitles attached to p_S, and used to create the output STM file.*

The second phone decoder (used for our contrastive-1 and contrastive-3 systems) was trained on BP data extracted in a semi-supervised fashion from the audios and minutes of BP plenary sessions, as described in [7]. The resulting BP dataset amounts to more than 1000 hours of speech, again with a high imbalance of Spanish over Basque (this time with a 2:1 ratio).

To build the phone decoders, an off-the-shelf (close to state-of-the-art) end-to-end neural network-based ASR system is used: Facebook AI Research wav2letter++ (consolidated into Flashlight), applying the Gated ConvNet recipe presented in [12].

3.2. The Grapheme-to-Phoneme (G2P) converter

An in-house bilingual Basque-Spanish rule- and dictionary-based grapheme-to-phoneme (G2P) converter is applied to get the phonetic baseforms of words in the subtitle texts, using the same reduced set of 23 phone-like units of our bilingual phone decoders. Before actually applying the G2P converter, the subtitle texts are normalized by deleting punctuation marks and expanding some known abbreviations as the most likely word in Spanish ('m.' expands as 'metros', 'km.' as 'kilometros', 'º' as 'grados', 'l.' as 'litros', '%' as 'por ciento', etc.). Numbers and ordinals are converted into their alphabetical counterparts using their most common realization (which might not match their actual pronunciation). Acronyms are transcribed as letter spellings, unless a specific pronunciation is found in the dictionary. When doing G2P conversion, the source words and subtitles are linked to their phonetic transcriptions. In this way, the timing information obtained from the alignment can be transferred back to those words and subtitles.

3.3. Modified dynamic programming kernel

The dynamic programming algorithm used to align the recognized and reference sequences p_X and p_S operates with a kernel (v_m, v_n, v_d, v_i), where v_m stands for the *value* added to the alignment path whenever a pair of matching units are found, and v_n, v_d and v_i stand for the values added to the alignment path by

the three types of alignment errors: non-matching units, deletions and insertions, respectively. Here, we implicitly consider that the added values are unit-independent. In a more general setup, those values might be different depending on the units being considered.

The baseline kernel used in this work is $(1, 0, 0, 0)$, which maximizes the number of matchings in the alignment path, no matter the number of errors. This kernel does not take into account the locations of matchings, and thus may lead to an optimal alignment path with matchings occurring far apart from one another (with any number of errors in-between). This would never happen (or very rarely) if the reference sequence (the subtitles) covered exhaustively the audio contents. But the subtitles used in this evaluation are not exahustive. They cover just some parts of the audio, the remaining parts being kind of *holes* that represent oportunities for badly located matchings. This is why sometimes our alignment algorithm hypothesizes very long words or very long subtitles, with some initial or final words being detected far ahead or far behind their actual locations, and a large number of insertions in-between, accounting for speech parts not included in the subtitles. It must be noted that this undesirable behaviour is not that pervasive to seriously harm the MAE metric, but it does introduce large errors at some points (around the *holes*) that increase the mean alignment error.

To fix this issue, we tried a kernel modification aiming to promote insertions between subtitles over insertions within subtitles: (1) a pseudo-unit '#' was inserted between every two subtitles in the reference sequence; and (2) the kernel was redefined as $(v_m, 0, 0, v_i^{\#})$ if the reference unit being considered was '#' and $(v_m, 0, 0, 0)$ if not. We explored several values for v_m and $v_i^{\#}$. The ones yielding the best performance on the development set were $v_m = 10$ and $v_i^{\#} = 2$. As a result, those alignment paths that had been optimal despite having some of their constituent words far apart from each other were no longer optimal, because other alignments with their constituent words close together were pushed over them, thanks to the new added value of between-subtitle insertions.

4. Results

Tables 1 and 2 show the alignment performance of GTTS systems on the dev and test sets, respectively. Late submissions (Con-2 and Con-3) are shown along with the originally submitted systems (Primary and Con-1). Our analysis and decisions were made based on the development set. This is why we present our results in two separate tables. Results in Table 1 suggest that the acoustic models trained on generic databases (Primary system) perform better than those trained on BP materials (Con-1), suggesting that, despite being three times larger than the generic acoustic database used for the primary system, the BP training dataset does not suitably match the TV broadcasts used in this evaluation. If we compare the performance of systems Con-2 (trained on generic data) and Con-3 (trained on BP data), the same conclusion can be drawn.

A second observation is that the modified kernel introduced in the late submitted systems (Con-2 and Con-3) is really making a difference. Though the average MAE/PTEM remains the same, its standard deviation gets lower than that of the systems using the baseline kernel (Primary and Con-1), suggesting that performance variability is reduced. And most importantly, the mean alignment error decreases remarkably, from 1.2665 to 0.8233, and from 1.1493 to 0.7950, meaning 35% and 30% relative reductions, respectively.

Table 1: *Performance of GTTS primary and contrastive systems on the dev set of TaSAC-ST (4 programs, 1894 subtitles). The minimum, maximum, average and standard deviation values of MAE/PTEM are shown, as well as the global mean of subtitle alignment errors.*

System	MAE/PTEM (sec)				Mean (sec)
	min	max	average	std-dev	
Primary	0.20	0.36	0.2950	0.0585	1.2665
Con-1	0.24	0.40	0.3250	0.0568	1.1493
Con-2	0.26	0.33	0.2950	0.0269	0.8233
Con-3	0.31	0.35	0.3250	0.0166	0.7950

Table 2: *Performance of GTTS primary and contrastive systems on the test set of TaSAC-ST (22 programs, 10600 subtitles). The minimum, maximum, average and standard deviation values of MAE/PTEM are shown, as well as the global mean of subtitle alignment errors.*

System	MAE/PTEM (sec)				Mean (sec)
	min	max	average	std-dev	
Primary	0.22	1.30	0.3950	0.2363	4.0923
Con-1	0.26	0.96	0.3986	0.1702	4.1990
Con-2	0.22	0.41	0.2927	0.0595	0.6053
Con-3	0.25	0.47	0.3277	0.0639	0.7186

The originally submitted systems (Primary and Con-1) perform much worse on the test set than on the dev set: the average MAE/PTEM increases from 0.2950 to 0.3950 for the primary system and from 0.3250 to 0.3986 for the Con-1 system, meaning 33.8% and 22.6% relative error increases, respectively. Performance variability gets higher too, the standard deviations getting multiplied by a factor of almost 4. But worst of all, the mean of the alignment errors also gets multiplied by almost 4, meaning that large alignment errors are being made. Fortunately, this issue gets fixed when using the modified kernel: the MAE/PTEM performance of Con-2 and Con-3 is almost the same on the test set than obtained by these systems on the dev set, meaning that they show a very robust behaviour across all kind of programs. Moreover, the mean of the alignment errors gets drastically reduced, from 4.0923 for the Primary system to 0.6053 for Con-2 and from 4.1990 for Con-1 to 0.7186 for Con-3, meaning 85% and 83% relative reductions, respectively.

To further illustrate the importance of the modified kernel introduced in our late submissions, Figure 2 shows the histograms of the alignment errors obtained by the Primary and Con-2 systems on the test set. Note that frequencies are shown in a logarithmic scale, to better appreciate the differences between the two systems. It can be clearly observed the huge difference that the new kernel makes, with Con-2 showing errors lower than 51 seconds in all cases, while the primary system produced a sizeable amount of alignment errors above that figure and up to 1540 seconds.

Finally, we aimed to study performance variability with regard to the TV programs used in the evaluation. Some programs might be tough to handle, such as, for example, interviews on the street or in crowded places, conversations with background music, noise or speech overlaps, etc. So we merged the alignments done by Con-2 (the most robust and accurate of our sys-

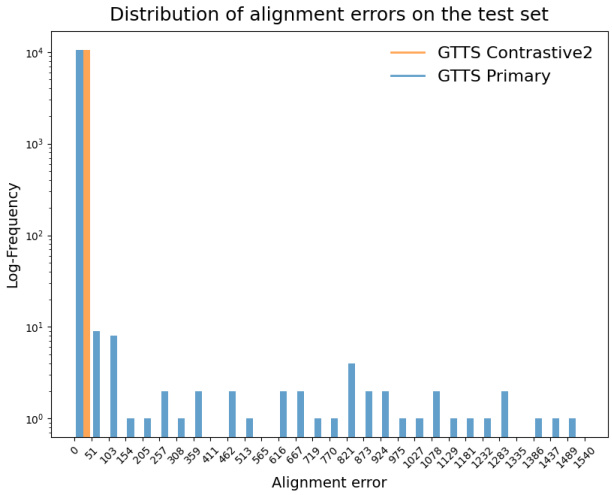

Figure 2: *Histograms of subtitle alignment errors obtained by the GTTS primary and contrastive-2 systems on the test set.*

tems), and disaggregated performance results by TV program (as shown in Table 3). It can be observed that MAE/PTEM performance and even the mean of the alignment errors are quite similar for AG (Agrosfera) and CO (Corazón), while performance is worse for AT (Aquí la Tierra). Differences are not large in terms of MAE/PTEM but remarkable in terms of the mean error, which may indicate that the alignment task could be hard in the case of challenging audios and/or poor subtitles.

Table 3: *Performance of the GTTS contrastive-2 system on the three programs used in the dev and test sets of TaSAC-ST: AG (Agrosfera, 10 programs), AT (Aquí la Tierra, 6 programs) and CO (Corazón, 10 programs). The minimum, maximum, average and standard deviation values of MAE/PTEM are shown, as well as the global mean of subtitle alignment errors.*

Program	MAE/PTEM (sec)				Mean (sec)
	min	max	average	std-dev	
AG	0.23	0.40	0.2850	0.0557	0.5250
AT	0.27	0.35	0.3100	0.0258	0.9177
CO	0.22	0.41	0.2910	0.0655	0.6112

4.1. Computational resources

Our phone decoder was run on a 2 x Intel Xeon CPU ES-2630 v3 @2.4 GHz, with 32 cores, RAM of 132 GB and an NVIDIA Titan X GPU. It took 9 minutes and 3 seconds to decode the 22 audio files of the test set. The remaining tasks (text normalization, G2P conversion, alignment of phone sequences using the baseline kernel, and postprocessing) were run on a desktop computer, an Intel Core-i9-19200 @2.4 GHz. It took 3 minutes and 6 seconds to perform those tasks for the test set. Finally, the alignments based on the modified kernel were run in a single thread on a 2 x Intel Xeon CPU ES-2450 @2.1 GHz, with 32 cores and RAM of 164 GB. It took 25 seconds and 2.5 GB of peak memory to perform the modified kernel alignments of the test set.

5. Conclusions

In this paper, the main features of the systems developed by GTTS for the Albayzin 2022 Speech and Text Alignment Challenge have been described and the results obtained have been presented and briefly discussed. Our approach relies on a phone decoder and a grapheme-to-phoneme converter which allow us to align the recognized and reference phone sequences and transfer timing information from the former to the latter. Two different sets of non-RTVE data have been used to estimate the acoustic models, which may be hindering the ability of our systems to decode the audio files used in this evaluation. A modified kernel has been introduced in our dynamic programming algorithm to promote insertions between subtitles over insertions within subtitles, so that more compact alignments are obtained and large alignment errors (such as the ones found with the baseline kernel) are avoided. Future work includes using RTVE data to estimate our acoustic models, carrying out a more in-depth study of the alignment errors and exploring further refinements of our dynamic programming kernel.

6. Acknowledgements

This work was partially funded by the Spanish Ministry of Science and Innovation (OPEN-SPEECH project, PID2019-106424RB-I00) and by the Basque Government (IT-1355-19).

7. References

[1] *Albayzin Evaluations - IberSpeech 2022 Evaluation Challenges*, Spanish Thematic Network on Speech Technologies (RTTH) and Cátedra RTVE - Universidad de Zaragoza, 2022, [Link].

[2] E. Lleida, A. Ortega, A. Miguel, V. Bazán, C. Pérez, and A. de Prada, *Albayzin Evaluation: IberSPEECH-RTVE 2022 Text and Speech Alignment Challenge: Alignment of re-spoken subtitles - TaSAC-ST Evaluation Plan*, Vivolab, Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain; Corporación Radiotelevisión Española, Spain, 2022, [Link].

[3] G. Bordel, M. Peñagarikano, L. J. Rodriguez-Fuentes, and A. Varona, *Albayzin 2022 Text-to-Speech Alignment System Evaluation: Subtask 2 - Evaluation Plan*, Grupo de Trabajo en Tecnologías Software (GTTS), UPV/EHU, 2022, [Link].

[4] G. Bordel, S. Nieto, M. Penagarikano, L. J. Rodriguez-Fuentes, and A. Varona, "Automatic Subtitling of the Basque Parliament Plenary Sessions Videos," in *Interspeech 2011*, Florence, Italy, 28-31 August 2011, Link.

[5] G. Bordel, M. Penagarikano, L. J. Rodriguez-Fuentes, and A. Varona, "Aligning very long speech signals to bilingual transcriptions of parliamentary sessions," in *Iberspeech 2012*, Madrid, Spain, November 21-23 2012, Link.

[6] G. Bordel, M. Penagarikano, L. J. Rodriguez-Fuentes, A. Álvarez, and A. Varona, "Probabilistic kernels for improved text-to-speech alignment in long audio tracks," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 126–129, January 2016, Link.

[7] M. Penagarikano, A. Varona, G. Bordel, and L. J. Rodriguez-Fuentes, "Semisupervised training of a fully bilingual ASR system for Basque and Spanish," in *Proceedings of IberSpeech, Granada (Spain), 14-16 November*, 2022.

[8] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," *CoRR*, vol. abs/1912.06670, 2019. [Online]. Available: http://arxiv.org/abs/1912.06670

[9] O. Kjartansson, A. Gutkin, A. Butryna, I. Demirsahin, and C. Rivera, "Open-Source High Quality Speech Datasets for Basque, Catalan and Galician," in *Proceedings of the 1st Joint Workshop on SLTU and CCURL*, Marseille, France, May 2020, pp. 21–27, Link.

[10] I. Odriozola, I. Hernaez, M. Torres, L. J. Rodriguez-Fuentes, M. Penagarikano, and E. Navas, "Basque Speecon-like and Basque SpeechDat MDB-600: Speech Databases for the Development of ASR Technology for Basque," in *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 26-31 2014, pp. 2658–2665, Link.

[11] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J. B. Marino, and C. Nadeu, "Albayzin speech database: design of the phonetic corpus," in *Proc. 3rd European Conference on Speech Communication and Technology (Eurospeech 1993)*, 1993, pp. 175–178, Link.

[12] R. Collobert, C. Puhrsch, and G. Synnaeve, "Wav2Letter: an End-to-End ConvNet-based Speech Recognition System," *CoRR*, vol. abs/1609.03193, 2016. [Online]. Available: http://arxiv.org/abs/1609.03193