Improving a Long Audio Aligner through Phone-relatedness Matrices for English, Spanish and Basque

Aitor Álvarez, Pablo Ruiz, and Haritz Arzelus

Human Speech and Language Technologies, Vicomtech-IK4, San Sebastián, Spain
{aalvarez,pruiz,harzelus}@vicomtech.org

Abstract. A multilingual long audio alignment system is presented in the automatic subtitling domain, supporting English, Spanish and Basque. Pre-recorded contents are recognized at phoneme level through language-dependent triphone-based decoders. In addition, the transcripts are phonetically translated using grapheme-to-phoneme transcriptors. An optimized version of Hirschberg's algorithm performs an alignment between both phoneme sequences to find matches. The correctly aligned phonemes and their time-codes obtained in the recognition step are used as the reference to obtain near-perfectly aligned subtitles. The performance of the alignment algorithm is evaluated using different non-binary scoring matrices based on phone confusion-pairs from each decoder, on phonological similarity and on human perception errors. This system is an evolution of our previous successful system for long audio alignment.

Keywords: Long audio alignment, automatic subtitling, phonological similarity matrices, perceptual confusion matrices.

1 Introduction

Subtitling is one of the most important means to make audiovisual content accessible. To promote accessibility, current European audiovisual law is forcing TV channels to subtitle a huge proportion of their contents. To address this increased demand, broadcasters and subtitlers are seeking alternatives more productive than manual subtitling. Speech recognition technologies have proved useful in this respect. One efficient approach, when the script for the content exists, is speech-text alignment, which relies on aligning audio with its script to automatically recover time stamps. Forced-alignment is challenging with long signals, because of the widely-used Viterbi algorithm, which forms very large lattices during decoding, requiring a lot of memory.

In this work, the system presented in [1] for long audio alignment in an automatic subtitling scenario has been improved and extended. Phone-decoder accuracy was improved using context-dependent acoustic models, besides implementing an adaptation of the generic language models to the script of the contents to subtitle. The system was also extended to Basque, its original languages being English and Spanish, and additional linguistic resources were created for the Spanish aligner.

The paper is structured as follows. Section 2 looks at related work in long audio alignment and in phone-relatedness measures. Section 3 describes our speech-text alignment system, and Section 4 presents the phoneme similarity matrices created. Section 5 discusses the evaluation method and results. Section 6 presents conclusions and suggestions for further work.

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 473–480, 2014.
This is preprint prepared by Proceedings editor for Springer International Publishing Switzerland.

2 Related Work

The reference for many of the related studies is the work done in [2], where the forced alignment was turned into a recursive and iteratively adapted speech recognition process. They used dynamic programming to align the hypothesis text and the reference transcript at word level. Subsequent works proposed improvements of this system, to deal with scenarios in which transcripts are not exact. In [3] a Driven Decoding Algorithm (DDA) was proposed to simultaneously align and correct the imperfect transcripts. At a new generated assumption of the speech recognizer in the lattice, DDA aligned it with the approximated transcript and a new matching score was computed and integrated with the language model for linguistic rescoring. An efficient, and simpler, long audio alignment approach was presented in [4]. They developed a system based on Hirschberg's dynamic programming algorithm [5] to align the phone decoder output with the transcription at phoneme level. They used a binary matrix to score alignment operations, with a cost of one for insertions, deletions and substitutions, and a cost of 0 for matches. Inspired on [4], for our experiments in [1] we created several scoring matrices, based on criteria like phonological similarity, phone-decoder confusion and phone confusion in human perception.

Concerning literature relevant for the creation of our scoring matrices, our phonological similarity metric is based on [6], where Kondrak constructed a metric that outperformed previously available ones, evaluating it with cognate alignment tasks. The metric was also successfully employed in spoken document retrieval in [7]. Regarding phone confusion in human perception, our American English matrices rely on perceptual error data reported in [8], who used a phoneset that closely corresponds to our phone-decoder's phoneset. Our Spanish data are based on the corpus of misperceptions developed by [9], which provides data covering our entire phoneset.

3 Long Speech-text Alignment System

The goal of any speech-text alignment system is to obtain a perfect timing synchronization between the source audio and related text recovering the time codes for each word in the transcript. Our multilingual long speech-text alignment system is trained to align long audios and related transcripts for English, Spanish and Basque. For each language, a language-dependent phone decoder was developed, in addition to a grapheme-to-phoneme transcriptor. The aim of the alignment algorithm is to find matches between the phones recognized by the phone-decoder and the reference phoneme transcription. Only the time-codes of the correctly aligned phones will be used as reference times for further synchronization.

However, all the phonemes are not always correctly aligned during alignment; substitutions, deletions and insertions may occur. In fact, using the evaluation contents presented in Section 5, only 34% of the phonemes were correctly aligned for English, while 57% and 48% of the phonemes were matched for Spanish and Basque respectively in the best-performing configuration. These time-codes at phoneme level are then used to estimate the start time of each word and thus of each subtitle. The promising results presented in this paper prove that the time-codes recovered by the aligner are good enough to generate near-perfectly aligned subtitles.

3.1 Context-dependent Phone Decoders

The phone-decoders have been improved from the last version of the system presented in [1], in which monophone models were employed. For this study, cross-word triphone models were built for each language to deal with coarticulation effects. With the aim of reducing linguistic variability, the language model consisted of an interpolation of the generic language model and a specific model created for each transcript. The interpolated models were bigram triphone models. The triphone-based phone decoders were trained using the HTK[1] tool. The parametrization of the signal consisted of 18 Mel-Frequency Cepstral Coefficients plus the energy and their delta and delta-delta coefficients, using 16-bit PCM audios sampled at 16 KHz.

The English triphone-based decoder system was built using the TIMIT database [11], which is composed by 5 hours and 23 minutes of clean speech data. Texts totaling 369 million words, gathered from digital newspapers, were used to train the generic language model. The Phone Error Rate (PER) of this decoder was 24.71%.

The Spanish triphone-based decoder system was based on 20 hours of clean-speech from three databases; Albayzin [12], Multext [13], and records of broadcast news contents from the SAVAS corpus [14]. The generic language model was trained with texts crawled from national newspapers, toting up 45 million words. The PER of the Spanish decoder was 31.79%.

The Basque triphone-based decoder system was generated using 36 hours of clean speech records of broadcast news contents. The generic language model was built using texts crawled from national newspaper, totaling 91 million words. The PER of the Basque decoder was 20.92%.

For all three languages, the corpora were split between training and test sets containing 70% and 30% of the data, respectively.

3.2 Grapheme-to-Phoneme Transcriptors

The grapheme-to-phoneme (G2P) transcriptors used for English and Spanish were the same used in the previous work [1]. The Spanish G2P was ruled based and inspired on the tool provided by Lopez[2]. The English transcriptor was inferred from the Carnegie Mellon Pronouncing Dictionary[3] using Phonetisaurus[4] tool. The Basque G2P transcriptor was based on manually created heuristic rules. The phonesets for all the languages are available on our project's website[5].

3.3 Algorithm for Alignment of Phoneme Sequences

Our alignment algorithm is a slightly modified version of the well-known divide-and-conquer Hirschberg's algorithm. These modifications were established once their effectiveness in the alignment process was tested.

[1] http://htk.eng.cam.ac.uk/

[2] http://www.aucel.com/pln/

[3] http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/

[4] http://code.google.com/p/phonetisaurus/

[5] http://sites.google.com/site/similaritymatrices/

Given the two phoneme sequences $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_m\}$ to be aligned, the algorithm forces them to be recursively divided at indexes x_{mid} and y_{mid} respectively. Hirschberg defined x_{mid} as *round(length(x)/2)*. Nevertheless, following the procedure several candidates can arise for y_{mid}. In our algorithm, y_{mid} always corresponds to the candidate-index closest to the middle of Y. The other modification relies on forcing a substitution operation, even if the phonemes do not match, when the recursive algorithm only has sequences of one symbol left to align.

Four edit-operations are allowed in the alignment algorithm: matches, substitutions, deletions and insertions. The scores for matches and substitutions are defined by the scoring matrices (See Section 4), while deletions and insertions incur a gap penalty. Since each matrix-type tested has a different range of values, the gap penalties are also different for each matrix-type. In our binary matrix, the gap penalty was 2. For all other matrices, the penalty was a quarter of the matrix' maximum value, following one of the practices for gap penalties referenced in [6].

4 Phoneme-relatedness Scoring Matrices

The phoneme-relatedness matrices provide information to the aligner about how likely it is for an alignment between two phonemes to be correct. The matrices favour aligning similar phonemes, by giving such alignments higher scores than to alignments between less similar phonemes. The matrices give the lowest scores to alignments between highly dissimilar phones, which are unlikely to be correct.

We created different scoring matrices for each language, applying different phoneme-relatedness criteria. The first scoring matrix is decoder-dependent, based on errors made by the phone decoder. The second matrix is decoder-independent, and based on phonological similarity, assessed by comparing largely articulatory features. The final matrix is also decoder-independent, and relies on phoneme confusion in human perception. Samples for all types of matrices are available on our project's website.

4.1 Matrices based on Phone-decoding Errors

The matrices were created based on HTK's HResults logs, when aligning the phone-decoding output and the G2P transcription for sequences of approx. 200,000 phonemes in English, 1,000,000 in Spanish and 2,000,000 in Basque. For each phone in the phone-set, the matrices contain the percentages of misrecognitions and correct recognitions by the decoder, normalized to a 1–1000 integer range. For instance, if 4% of the occurences of /ɲ/ were misrecognized as /n/, the matrix shows a score of 40 for the [ɲ,n] phoneme pair. In order to prevent substitutions between phonemes never mistaken by the decoder, a score of -500 was entered in the matrix for such phoneme-pairs. This score corresponds to ½×(0 – max({Score Range}).

4.2 Matrices based on Phonological Similarity

Our phonological similarity scores are based on the metric devised by Kondrak in [6], as part of the ALINE cognate alignment system[6]. Phonemes are described with Lade-

[6] ALINE is available at `http://webdocs.cs.ualberta.ca/~kondrak/#Resources`

foged's [14] multivalued features, and a *salience* factor weights each feature according to its impact for phoneme similarity. The features, values and saliences employed for each language are available on our project's website.

$$\sigma_{\text{sub}}(p,q) = (C_{\text{sub}} - \delta(p,q) - V(p) - V(q))/100$$

$$\text{where } V(p) = \begin{cases} 0 & \text{if } p \text{ is a consonant or } p = q \\ C_{\text{vwl}} & \text{otherwise} \end{cases}$$

$$\delta(p,q) = \sum_{f \in R} \mathit{diff}(p,q,f) \times \mathit{salience}(f)$$

$$\sigma_{skip}(p) = \mathit{ceiling}(|C_{\text{sub}}/400|)$$

Fig. 1. Similarity function, based on Kondrak (2002)

Fig. 1 shows equations with our scoring function. $\sigma_{\text{sub}}(p, q)$ returns the similarity score for phonemes p and q, $C_{\text{sub}}/100$ being the maximum possible similarity score. C_{vwl} represents the relative weight of consonants and vowels. Values for C_{sub} and C_{vwl} are set heuristically as described in [15]. The function diff(p, q, f) yields the similarity score between phonemes p and q for feature f, and the feature-set R is configurable. Last, $\sigma_{\text{skip}}(p)$ returns the penalty for insertions and deletions used in the aligner. We defined heuristically a C_{sub} value of 3,500 (i.e. a maximum similarity score of 35), and a gap penalty of 9 for alignment, which corresponds to ceiling($|\text{C}_{\text{sub}}/400|$).

Kondrak's original function was designed for cognate alignment. We modified the function, for coherence with our audio aligner, and to adapt it to audio alignment tasks, achieving better results with the modified version than with the original. Details about the modifications are discussed in [1] and in the project's website.

4.3 Matrices Based on Perceptual Errors

The English matrices were based on human perceptual error data from [8]. They performed a phoneme identification study with native speakers of American English, asking them to identify the initial or final phoneme of 645 syllables of types CV (ConsonantVowel) and VC, at signal-to-noise ratios (SNR) of 0, 8 and 16. The noise type was multi-speaker babble. Participants chose a response among several possibilities presented to them visually. The phoneme-set in the study covers all of our decoder's phoneset except schwa. We only used the SNR 16 data, since a matrix based exclusively on this subset of the data yielded better alignment results than when considering data at other SNR for building the matrix.

The Spanish matrix was based on an extended version, provided by the authors directly, of the corpus of human misperceptions in noise developed in [9]. The methodology involved presenting 69 native speakers of Spanish with over 20,000 single-word stimuli, under different masking-noise conditions, and asking the speakers to write the word they had heard. Only stimuli for which certain agreement thresholds were reached among participants' responses were kept for the final misperception corpus, which consists of 3,294 stimuli and their associated responses. The study is thus a free-response error-elicitation task, not a closed-response task like [8]. However, we chose [9] as our data source, since, unlike other Spanish perception studies, it provides data for all phonemes in our decoder's phoneset. For coherence with our English data, we based our matrices on the 1,838 stimuli where multi-speaker babble was used as the masker. SNR in these stimuli ranged between −8 and +1. For computing our confusion matrix, we compared the corpus' stimulus and responses in cases where the response involved a single-phoneme error. We recorded the percentage of matches and mismatches between each stimulus and each response in the stimulus' response-set (a maximum of 15 responses were available per stimulus). Match and mismatch percentages were normalized to a 1–1000 range. For phoneme pairs where no confusion had taken place, a score of −500 (i.e. *½×(0 – max({Score Range}*) was entered in the matrix. The matrix was based on 6,807 stimulus-response pairings.

Perceptual-relatedness matrices were not created for Basque, since we are not aware of appropriate data that could be exploited for their creation.

5 Evaluation and Results

The English test-set totaled 21,310 phonemes, 4,732 words and 471 subtitles, and contained non-clean speech from television audios. Its reference subtitles contained some stretches where transcription was imperfect, with subtitles missing for some parts of the audio. The Spanish test-set consisted of 47,480 phonemes, 8,774 words and 1,249 subtitles, and was composed of clean speech from documentaries. The Basque test-set totaled 26,712 phonemes, 4,331 words and 726 subtitles, containing a concatenation of a documentary and a film, and included noisy-speech.

Table 1. Alignment accuracy at word and subtitle level. **PDE**: Phone-decoder-error based matrix, **PHS**: Phonological similarity, **PCE**: Perceptual error matrix.

	Word-level deviation (seconds)						**Subtitle-level deviation (seconds)**					
	Matrix	**0**	**≤0.1**	**≤0.5**	**≤1.0**	**≤2.0**	**0**	**≤0.1**	**≤0.5**	**≤1.0**	**≤2.0**	**Matrix**
English	Binary	0.25	8.13	28.12	40.36	56.14	0.42	4.46	38.64	83.65	100	Binary
	PDE	**1.02**	**29.88**	**60.17**	**72.94**	**84.52**	**0.85**	**14.86**	**54.35**	**91.30**	100	PDE
	PHS	0.87	25.41	56.10	69.55	79.73	0.64	11.25	53.50	88.54	100	PHS
	PCE	0.72	26.66	57.26	70.31	82.57	0.64	14.65	53.29	90.02	100	PCE
Spanish	Binary	2.47	47.70	69.11	75.55	80.21	0.48	21.06	63.49	92.47	100	Binary
	PDE	**5.55**	77.42	**92.21**	**94.39**	95.93	1.12	40.19	**80.22**	96.64	100	PDE
	PHS	5.44	**77.45**	92.17	94.31	**95.97**	**1.20**	**40.67**	80.14	96.64	100	PHS
	PCE	5.22	74.83	92.03	94.48	96.35	1.28	38.83	78.78	**96.72**	100	PCE
Basque	Binary	1.55	34.98	56.63	61.06	65.25	0.83	24.24	64.05	92.29	100	Binary
	PDE	2.34	48.91	76.21	80.91	85.05	**1.65**	**44.63**	**75.76**	**95.18**	100	PDE
	PHS	**2.49**	**49.97**	**77.00**	**82.10**	**86.45**	1.38	35.54	74.24	95.04	100	PHS

Long audio alignment accuracies using different phone-relatedness matrices for English, Spanish and Basque are presented in Table 1. The results present the percentage or words and subtitles correctly aligned within the specified deviation range from the reference. The real time-codes at word level were obtained applying a forced-alignment algorithm for each subtitle in the reference material, which was composed of time-coded subtitles manually created by professional subtitlers. For subtitle-level evaluation, the deviation of the first and last words of the subtitles were measured.

The results show the effectiveness of our long audio alignment system, even with contents containing noisy-speech and imperfect transcriptions. Besides, the improvements using non-binary matrices are clearly proved comparing to the accuracies obtained with the binary matrix. Considering that a maximum deviation of 1 second is not long enough for listeners to have difficulties associating the subtitle and the audio, near-perfectly aligned subtitles were obtained for all three languages. In fact, alignment accuracies of 91.30%, 96.72% and 95.18% were obtained for English, Spanish and Basque respectively at this maximum deviation time.

Regarding non-binary matrices performance, the PDE matrices achieve the most accurate alignment results for English and Basque. It was expectable since these matrices were based on each phone-decoder phone confusion-pairs. However, the improvements with the PDE matrix comparing to improvements with the other non-binary matrices are not relevant. For Spanish, the PCE matrix obtained the best results, although the PDE and PHS matrices achieved very similar accuracies.

6 Conclusions and Further Work

The adequate performance of our multilingual long audio alignment system in the automatic subtitling scenario was presented in this work. We established the effectiveness of a customized version of the well-known Hirschberg algorithm, and proved that using several scoring matrices based on different phoneme-relatedness criteria obtains well-performed alignments.

Since the current system works with triphone-based phone decoders, ongoing work is focused on the development of context-dependent phoneme scoring matrices. The goal behind this approach will be to improve the alignment process considering not only phones, but also biphones and triphones, to deal with coarticulation effects.

References

1. Álvarez A., Arzelus, H., Ruiz P.: Long audio alignment for automatic subtitling using different phone-relatedness measures. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Florence, Italy (2014).
2. Moreno, P.J., Joerg, C., Van Thong, J-M and Glickman, O.: A recursive algorithm for the forced alignment of very long audio segments. In: Proceedings of the 5th International Conference on Spoken Language Processing, ICSLP, Sydney, Australia, (1998).
3. Lecouteux, B., Linàres, G., Nocéra, P., Bonastre, J.: Imperfect transcript driven speech recognition. In: Proceedings of INTERSPEECH, pp. 1626–1629 (2006).
4. Bordel, G., Nieto, S., Peñagarikano, M., Rodríguez-Fuentes L.J., Varona, A. A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions. In: Proceedings of INTERSPEECH, Portland, Oregon (2012).
5. Hirschberg, D.S.: A linear space algorithm for computing maximal common subsequences. Communications of the ACM, vol. 18, no. 6, pp. 341–343, (1975).

6. Kondrak, G.: Algorithms for Language Reconstruction, PhD Thesis. University of Toronto (2002).
7. Comas, P.: Factoid Question Answering for Spoken Documents, PhD Thesis. Universitat Politècnica de Catalunya (2012).
8. Cutler, A., Weber, A., Smits, R., Cooper, N.: Patterns of English phoneme confusions by native and non-native listeners. In: Journal of the Acoustical Society of America, vol. 116, no. 6, pp. 3668–3678 (2004).
9. García Lecumberri, M.L., Toth, A.M., Tang Y., Cooke. M.: Elicitation and analysis of a corpus of robust noise-induced word misperceptions in Spanish. In: Proceedings of INTERSPEECH, pp. 2807–2811 (2013).
10. Garafolo, J.S.L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren N., Zue, V.: TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium, Philadelphia (1993).
11. Díaz, J.E., Peinado, A., Rubio, A., Segarra, E., Prieto N., Casacuberta, F.: Albayzín: a task-oriented Spanish speech corpus. In: Proceedings of LREC, Granada, Spain (1998).
12. Campione E., Véronis, J.: A multilingual prosodic database. In: Proceedings of the 5th International Conference on Spoken Language Processing, ICSLP, Sydney, Australia (1998).
13. Del Pozo, A., Aliprandi, C., Álvarez, A., Mendes, C., Neto, J.P., Paulo, S., Piccinini, N., Rafaelli, M.: SAVAS: Collecting, Annotating and Sharing Audiovisual Language Resources for Automatic Subtitling. In: Proceedings of LREC, Reykjavik, Iceland (2014).
14. Ladefoged, P.: A Course in Phonetics, New York: Harcourt Brace Jovanovich (1995).
15. Ruiz, P., Álvarez, A., Arzelus, H.: Phoneme similarity matrices to improve long audio alignment for automatic subtitling. In: Proceedings of LREC, Reykjavik, Iceland, (2014).