Phoneme Similarity Matrices to Improve Long Audio Alignment for Automatic Subtitling

Pablo Ruiz, Aitor Álvarez, Haritz Arzelus

Phoneme Similarity Matrices to Improve Long Audio Alignment for Automatic Subtitling

Pablo Ruiz, Aitor Álvarez, Haritz Arzelus
Vicomtech-IK4
Mikeletegi Pasealekua, 57, 20009 Donostia/San Sebastián, Spain
E-mail: {pruiz,aalvarez,harzelus}@vicomtech.org

Abstract

Long audio alignment systems for Spanish and English are presented, within an automatic subtitling application. Language-specific phone decoders automatically recognize audio contents at phoneme level. At the same time, language-dependent grapheme-to-phoneme modules perform a transcription of the script for the audio. A dynamic programming algorithm (Hirschberg's algorithm) finds matches between the phonemes automatically recognized by the phone decoder and the phonemes in the script's transcription. Alignment accuracy is evaluated when scoring alignment operations with a baseline binary matrix, and when scoring alignment operations with several continuous-score matrices, based on phoneme similarity as assessed through comparing multivalued phonological features. Alignment accuracy results are reported at phoneme, word and subtitle level. Alignment accuracy when using the continuous scoring matrices based on phonological similarity was clearly higher than when using the baseline binary matrix.

Keywords: phoneme similarity matrices, long audio alignment, automatic subtitling

1. Introduction

Accessibility needs, and policies addressing them, are stimulating a large demand for subtitling in the broadcast industry. Manual subtitling being time and labour-intensive, automatic subtitling is an attractive option, as it saves time and resources.

Our approach to automatic subtitling aligns the audio signal with a human transcript for the audio. Aligning long audio signals is challenging, given memory demands, processing time and error-proneness of algorithms when aligning long sequences.

A successful system for long audio alignment is Bordel et al. (2012). They report alignment results for 3-hour long audios. Their alignment method is based on Hirschberg's algorithm (1975), originally used for genetic sequence alignment. The scoring matrix for alignment operations in Bordel et al. is binary: insertions, deletions and substitutions bear a cost of 1, and matches bear a cost of 0. In this paper, we follow Bordel et al.'s long audio alignment approach, improving one aspect: We show that, as compared with results for a binary matrix, scoring alignment operations with a matrix based on phoneme-similarity improves alignment results at phoneme level, word level and subtitle level. We present results for the alignment of long audios in Spanish and English.

Our similarity scores follow Kondrak's metric (2002), based on multivalued phonological features weighted by salience. The metric has been successfully employed in cognate alignment and spoken document retrieval (Comas, 2012).

Other phoneme similarity metrics based on phonetic or phonological criteria have been proposed for use in speech technology applications, e.g. Melnar and Liu (2006). We adopted Kondrak's metric for our phone similarity scoring since previous successful applications have been documented, and given ease of implementation.

The paper is structured as follows. Section 2 presents our long audio alignment system, and Section 3 describes the similarity matrices created. Section 4 discusses evaluation methods and results. Section 5 contains conclusions and suggestions for future work.

2. Speech-text alignment system

The speech-text alignment system aligns two sequences of phonemes obtained from different sources. Given the audio and the transcript of the content to be automatically subtitled, a language-dependent phone decoding is used to recognize phonemes and their time-codes from the audio. In addition, a grapheme-to-phoneme module translates the input transcript into the reference phoneme transcription. An alignment algorithm finds phoneme correspondences between the reference phoneme transcription and the phonemes recognized by the phone-decoder, which usually contain common recognition errors. Aligned phonemes are assigned the time-codes obtained by the phone-decoder. Phoneme alignment may present substitutions, deletions and insertion errors. However, the number of phone correspondences found generally provides enough time-codes to create subtitles with near-perfect alignment at word-level.

2.1. Phone decoding module

The phone decoding module was trained using HTK[1], a hidden Markov model toolkit. The acoustic model was based on a monophone model, with three left-to-right emitting states using 32 Gaussian mixture components. The language model was a bigram phoneme model. 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 Spanish phone-decoder was trained and tested with 20 hours of audios from three databases; Albayzín (Díaz

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

et al., 1998), Multext (Campione and Véronis, 1998) and records of clean-speech broadcast news from the Spanish subset of the SAVAS[2] corpus (Del Pozo et al., 2014). The contents were mixed and divided into training (70%) and test (30%) sets. Texts totaling 45 million words were crawled from a national newspaper to train the language model. The Spanish phone-decoder yielded a Phone Error Rate (PER) of 40.65%.

The English phone-decoder was trained and tested on the TIMIT database (Garafolo et al., 1993), which consists of 5 hours and 23 minutes. 70% of the database was used for training, leaving the rest for testing. Texts totaling 369 million words, collected from digital newspapers, were used to train the language model. The English phone-decoder yielded a PER of 35.52%.

2.2. Grapheme-to-phoneme transcriptors

Grapheme-to-phoneme (G2P) transcriptors were developed for Spanish and English. The Spanish transcriptor was rule-based, inspired on an open-source tool [3], and adapted to our phonelist. The English transcriptor was inferred from the Carnegie Mellon Pronouncing Dictionary [4] (CMUdict) using Phonetisaurus [5], a G2P framework based on weighted finite state transducers (WFST).

The Spanish and English phonesets are available on our project's website.[6]

2.3. Algorithm for long sequence alignment

We used Hirschberg's (1975) algorithm, an optimization of Needleman and Wunsch's (1970) algorithm to calculate the optimal alignment of two sequences of length n and m in $n \times m$ steps.

Each alignment operation receives a score, and the alignment obtaining the best score is chosen. Substitutions are evaluated with a scoring matrix. Gaps (insertions and deletions) incur a penalty. When aligning with the binary scoring matrix, our gap penalty was 2. When using the phoneme-similarity based matrices, our gap penalty was 10, based on parameter C_{skip} from our similarity function (see Section 3).

Needleman-Wunsch has been successfully applied to many problems, but it requires a large amount of space with long sequences; Θ(nm) for strings of length n and m. Hirschberg developed a space reduction method based on the Needleman-Wunsch algorithm, decreasing the required space from Θ(nm) to Θ(min{n,m}) while only doubling the worst-case processing time.

Bordel et. al (2012) based their system on Hirschberg's algorithm, showing its suitability. Nevertheless, they used a binary scoring matrix, while in the present study matrices based on phoneme similarity were developed. This improved alignment accuracy *vs.* a binary matrix.

[2] http://www.fp7-savas.eu/
[3] http://www.aucel.com/pln/
[4] http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict
[5] http://code.google.com/p/phonetisaurus
[6] https://sites.google.com/site/similaritymatrices/

3. Phoneme similarity matrices

Our similarity scores are based on the metric in Kondrak's (2002) ALINE cognate alignment system.[7] Phonemes are described with Ladefoged's (1995) multivalued features. Features are weighted according to their *salience*: the feature's impact for similarity. Features *place* and *manner* need to bear significantly higher salience than the rest.

The phoneme and feature set, feature values and salience weights need to be adapted to each language. For each phone in our Spanish and English phonesets, we created feature specifications, available on our project's website (see footnote 6). Samples are shown in Table 5. Salience weights are in Table 6.

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

where

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

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

$$\sigma_{skip}(p) = |C_{skip} / 100|$$

Figure 1: Similarity function

The scoring function is in Figure 1: $\sigma_{sub}(p, q)$ returns the similarity score for segments p and q. $C_{sub}/100$ is the maximum similarity score attainable. C_{vwl} determines the relative weight of consonants and vowels. Values for C_{sub} and C_{vwl} are set heuristically. The function diff(p, q, f) outputs the difference between segments p and q for feature f. The set of features R is configurable. Finally, $\sigma_{skip}(p)$ returns $C_{skip}/100$, which is used to define the penalty for insertions and deletions employed in the aligner (see Section 2.3).

Kondrak's original function contains an additional clause, not shown in Figure 1, to evaluate two-to-one phoneme alignments. We did not use that clause since many-to-many alignments are not implemented in our aligner. Another modification in our version of the similarity function, compared to Kondrak's, is that we added a denominator of 100 to σ_{sub} and σ_{skip}. This allowed us to keep our similarity scores in the range reported by Kondrak, but deploying integer feature values instead of decimals, and so reducing memory use.

We created different matrices, varying the settings for elements (1) through (3) below. Table 1 and Table 2 show a summary of the settings for each matrix. Table 3 and Table 4 show matrix samples.

For all matrices, C_{sub} was 3500, yielding a maximum possible similarity score of 35 ($C_{sub}/100$). C_{skip} was −1000, yielding a gap penalty of 10 when aligning ($|C_{skip}/100|$).

[7] http://webdocs.cs.ualberta.ca/~kondrak/#Resources for Kondrak's ALINE. A Python implementation (PyAline) by Huff (2010) is at http://sourceforge.net/projects/pyaline/

***(1)* C_{vwl}: 0 vs. 1000**. A desirable outcome of setting $C_{vwl} > 0$ is that substitutions between vowels and consonants are more clearly penalized by the matrix, getting lower scores than when $C_{vwl} = 0$. However, with $C_{vwl} > 0$, vowel matches get a lower similarity score than consonant matches, decreasing the weight of vowels in alignment. This is useful for cognate alignment (Kondrak, 2002, p. 48). The question arises whether this is also beneficial when aligning decoded phonemes and a G2P output. We tested this by defining *V(p)* in the scoring function differently.

***(2)* V(p): original vs. alternative definition**. The alternative definition of *V(p)* in Figure 2 allows us to give equal scores to vowel matches and consonant matches, while still setting $C_{vwl} > 0$, and thus still obtaining the beneficial effect of penalizing consonant/vowel substitutions more than consonant/consonant ones.

With parameters p, q from $\sigma_{sub}(p, q)$

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

Figure 2: Alternative definition for *V(p)*

***(3)* Diphthongs: binary vs. continuous scores**. This applies only to our English matrices. Our English phoneset treats diphthongs as single phones, but in Kondrak they are two-phoneme sequences. To score diphthong substitutions with Kondrak's function, we assigned them features and values heuristically. For comparison, we created matrices where diphthong scores were binary (match *vs*. mismatch).

Matrix Name	C_{vwl}	Definition of *V(p)*
C_v0_VpO	0	original
C_v1K_VpO	1000	original
C_v1K_VpA	1000	alternative

Table 1: Spanish Similarity Matrices and their settings

Matrix Name	C_{vwl}	Definition of *V(p)*	Diphthong Scores
C_v0_VpO_DB	0	original	binary
C_v0_VpO_DC	0	original	continuous
C_v1K_VpO_DB	1000	original	binary
C_v1K_VpA_DB	1000	alternative	binary
C_v1K_VpO_DC	1000	original	continuous
C_v1K_VpA_DC	1000	alternative	continuous

Table 2: English Similarity Matrices and their settings

IPA	a	i	n	p	r	s	j
a	35	7	−50	−56	−30	−50	2
i	7	35	−26	−32	−6	−26	10
n	−50	−26	35	9	−5	−5	−21
p	−56	−32	9	35	−11	9	−27
r	−30	−6	5	−11	35	−5	9
s	−50	−26	5	9	−5	35	−21
j	2	10	−21	−27	9	−21	35

Table 3: Sample from Spanish Matrix C_v*1K_VpA*

IPA	æ	i:	n	p	ɹ	s	aj
æ	35	9	−46	−57	−16	−36	10
i:	9	35	−26	−37	4	−16	−46
n	−46	−26	35	4	5	5	−46
p	−57	−37	4	35	−6	14	−46
ɹ	−16	4	5	−6	35	15	−46
s	−36	−16	5	14	15	35	−46
aj	10	−46	−46	−46	−46	−46	35

Table 4: Sample from English Matrix C_v*1K_VpA_DC*

4. Evaluation and results

We evaluated alignment at phoneme, word, and subtitle level, aligning long audios containing spontaneous speech, with disfluencies. The Spanish test-set was clean speech. The English test-set was non-clean speech, with music, noise and overlapping utterances. Accordingly, lower accuracy in English was expected and observed, at all evaluation levels. Another difficulty with English subtitles, which also led to lower accuracy, is that they represent a less literal transcription of the audio than the Spanish subtitles, due to a different subtitling approach in each language.

The test-sets are different to the ones used to evaluate the phone-decoder, and consist of television audios, providing results that are more indicative of alignment quality in a real application scenario.

The Spanish test-set contained 47,480 phonemes, 8,774 words and 1,249 subtitles. The English test-set contained 21,310 phonemes, 4,732 words and 471 subtitles.

4.1 Evaluation at phoneme level

The number of correctly aligned phonemes, based on the number of matches during the alignment process, increased when using the phoneme-similarity based matrices. Improvements were around 11 percentage points in Spanish, from 38.14% with the binary matrix to 49.69% with the best-performing phoneme-similarity based matrix. Improvements in English were around 12 percentage points (15.57% with the binary matrix vs. 27.91% with the best phoneme-similarity based matrix).

SPANISH																
IPA	**Vic**	**Place[1]**		**Manner[1]**		**V**	**Syl**	**Voi**	**Nas**	**Lat**	**Tri**	**High[1]**		**Back[1]**		**Ro[1]**	***E.g.***
a	**a**	velar	60	low vowel	0	1	100	100	0			low	0	front	100	0	v***a***
i	**i**	palatal	70	high vowel	40	1	100	100	0			high	100	front	100	0	d***i***
n	**n**	alveolar	85	stop	100	0	0	100	100	0	0						***n***o
p	**p**	bilabial	100	stop	100	0	0	0	0	0	0						***p***an
r	**R**	alveolar	85	approximant	60	0	0	100	0	0	100						pe***rr***o
s	**s**	alveolar	85	fricative	80	0	0	0	0	0	0						***s***on
j	**j**	palatal	70	high vowel	40	1	0	100	0	0	0	high	100	front	100	0	ho***y***
ENGLISH																	
IPA	**Vic**	**Place[1]**		**Manner[1]**		**V**	**Syl**	**Voi**	**Nas**	**Lat**	**Asp**	**High[1]**		**Back[1]**		**Ro[1]**	**Lo[1]**
æ	**ae**	palatal	70	low vowel	0	1	100	100	0			low	0	front	100	0	0
i:	**iy**	palatal	70	high vowel	40	1	100	100	0			high	100	front	100	0	100
n	**n**	alveolar	85	stop	100	0	0	100	100	0	0						
p	**p**	bilabial	100	stop	100	0	0	0	0	0	100						
ɹ	**r**	alveolar	85	approximant	60	0	0	100	0	0	0						
s	**s**	alveolar	85	fricative	80	0	0	0	0	0	0						
aj	**ay**	palatal	70	low vowel+ high vowel	16	1	100	100	0			low+ high	40	central+ front	70	0	100

[1]To compare with each other phonemes where **V=1**, *Place* and *Manner* are replaced with *High*, *Back*, *Round*, and, if available, *Long*.

Shaded cells indicate features that are not used to define similarity for the segment in the language

Abbreviations	**V**: Vowel, **Syl**: Syllabic, **Voi**: Voice, **Nas**: Nasal, **Lat**: Lateral, **Asp**: Aspirated, **Tri**: Trill **Ro**: Round, **Lo**: Long, **Vic:** ASCII-based phone code

Table 5: Samples from the Phoneset, Features and Feature Values for Spanish and English

Place	40	Nasal	10	High	5
Manner	50	Lateral	10	Back	5
Syllabic	5	Aspirated	5	Round	5
Voice	10	Trill	10	Long	1

Table 6: Salience Weights for each feature

4.2 Evaluation at word level

We adopted Moreno et al.'s (1998) measure of word-level alignment, also used by Bordel et al. As Table 7 and Table 8 show, we record the cumulative percentage of correctly aligned words within a given deviation range: Column 0 shows the percentage of perfectly aligned words, column ≤0.1 means words whose misalignment goes up to 0.1sec, and so on. In the tables, we highlighted the best and worst results at 0, ≤0.1, ≤0.5 and ≤2 seconds.

Improvements with the phoneme-similarity based matrices were observed. In Spanish, exactly aligned words increased by ca. 9 percentage points, while improvement at a ≤0.5 deviation range was 20.85 percentage points. In English, improvements between ca. 20 and 30 percentage points were observed for each deviation range.

4.3 Evaluation at subtitle level

This is the most important evaluation, since it is indicative of the system's alignment quality in its application scenario: automatic subtitling. Reference subtitles were created manually by subtitling professionals.

For subtitle-level evaluation, we measured the deviation, compared to the reference, of the beginning of the subtitle's first word and of the end of the subtitle's last word. Cumulative percentages are given in Table 9 and Table 10.

In Spanish, when using the best-performing phoneme similarity based matrix, exactly aligned subtitles increased by 7.44 percentage points compared to results with the binary matrix. At the ≤0.5 deviation range, gains were 14.57 percentage points. In English, alignment improved at each deviation range, e.g. gains of 4.03 percentage points at ≤0.1 seconds and 8.92 percentage points at ≤0.5 seconds.

sec	0	≤0.1	≤0.3	≤0.5	≤1.0	≤1.5	≤2.0
Binary	**14.17**	**57.71**	70.09	**72.65**	76.21	78.04	**79.02**
C_v0_VpO	23.17	81.08	90.27	92.17	93.96	95.01	95.64
C_v1K_VpO	22.77	80.78	90.65	92.41	94.43	95.23	95.89
C_v1K_VpA	**23.01**	**82.21**	91.83	**93.50**	95.50	96.21	**96.83**

Table 7: **Spanish word alignment** accuracy. Percentage of words aligned within each deviation range from reference

sec	0	≤0.1	≤0.3	≤0.5	≤1.0	≤1.5	≤2.0
Binary	**10.57**	**45.08**	61.97	**73.26**	95.12	**99.76**	100
C_v0_VpO	18.33	65.73	82.23	87.35	98.56	**100**	100
C_v1K_VpO	17.93	65.57	82.95	87.43	98.56	99.84	100
C_v1K_VpA	**18.01**	**66.45**	82.71	**87.83**	98.80	99.84	100

Table 9: **Spanish subtitle alignment** accuracy. Percentage of subtitles aligned within each deviation range from reference

sec	0	≤0.1	≤0.3	≤0.5	≤1.0	≤1.5	≤2.0
Binary	**0.28**	**4.81**	12.73	**19.04**	29.69	37.38	**43.24**
C_v0_VpO_DB	1.59	19.57	34.94	45.00	58.22	65.46	70.31
C_v0_VpO_DC	1.59	19.86	34.33	42.78	56.61	63.64	68.72
C_v1K_VpO_DB	1.67	20.31	36.26	45.55	58.79	66.14	70.84
C_v1K_VpO_DC	1.67	20.03	33.91	42.23	54.76	61.10	64.85
C_v1K_VpA_DB	1.80	22.87	39.50	48.73	61.56	68.64	**73.08**
C_v1K_VpA_DC	**1.93**	**23.76**	40.03	**48.90**	61.58	68.34	72.72

Table 8: **English word alignment** accuracy. Percentage of words aligned within each deviation range from reference

sec	0	≤0.1	≤0.3	≤0.5	≤1.0	≤1.5	≤2.0
Binary	0.21	**4.25**	18.26	**37.15**	84.29	**98.73**	100
C_v0_VpO_DB	0.42	7.64	25.05	43.1	86.41	98.30	100
C_v0_VpO_DC	0.42	7.43	25.48	42.46	87.47	98.73	100
C_v1K_VpO_DB	0.42	8.92	26.54	40.55	87.47	98.51	100
C_v1K_VpO_DC	0.42	9.13	27.18	42.25	87.05	98.51	100
C_v1K_VpA_DB	0.42	**11.04**	30.79	45.86	87.47	98.94	100
C_v1K_VpA_DC	0.42	8.28	26.96	**46.07**	86.84	**99.58**	100

Table 10: **English subtitle alignment** accuracy. Percentage of subtitles aligned within each deviation range from reference

5. Conclusions and future work

This study shows that long audio alignment using Hirschberg's algorithm can be improved by using, instead of a binary scoring matrix, a scoring matrix based on phoneme similarity defined via phonological features. Improvements were observed at phoneme, word and subtitle level, when aligning both clean speech (Spanish tests) and non-clean speech (English tests).

As expectable, improvements at word level were higher than at subtitle level. At subtitle level, we only assess the position of the first and last word of each subtitle. This restricts the set of word-alignments that can contribute to a subtitle-level improvement.

Regarding the different matrices tested, we obtained slightly better results with the matrices created using a modified scoring function, that gives equal weight to consonant matches and vowel matches.

As future work, several approaches to improve alignment could be tested. First, our phoneme decoding applied MFCC coefficients, based on a perceptually motivated Mel frequency scale. However, our phoneme-similarity metric relied on phonological features that follow articulatory criteria. Using MFCC parametrization together with matrices based on perceptual similarity could be tested. The converse approach is also possible: Keeping a similarity metric based on articulatory criteria, but using an acoustic parametrization that provides a good description of the speech articulators, e.g. linear predictive coding (LPC). Finally, since alignment quality depends on phone-decoder accuracy, similarity matrices based on phone-decoding confusion matrices could be tested.

6. References

Bordel, G., S. Nieto, M. Peñagarikano, L. J. Rodríguez-Fuentes, A. Varona. (2012). A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions. In *Interspeech 2012, 13th Annual Conference of the International Speech Communication Association*. Portland, Oregon.

Campione, E. and J. Véronis. (1998). A multilingual prosodic database. In *ICSLP 1998*, *Proceedings of the 5th International Conference on Spoken Language Processing*. Sydney, Australia.

Comas, P. (2012). *Factoid Question Answering for Spoken Documents*. PhD Thesis. Universitat Politècnica de Catalunya.

Del Pozo, A, C. Aliprandi, A. Álvarez, C. Mendes, J. P. Neto, S. Paulo, N. Piccinini, M. Rafaelli. (2014). SAVAS: Collecting, Annotating and Sharing Audiovisual Language Resources for Automatic Subtitling. In *LREC 2014. Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik, Iceland.

Díaz, J. E., A. Peinado, A. Rubio, E, Segarra, N. Prieto and F. Casacubierta. (1998). Albayzín: a task-oriented Spanish speech corpus. In *LREC 1998. Proceedings of the First International Conference on Language*

Resources and Evaluation. Granada, Spain.

Garafolo, J. S, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, Philadelphia.

Hirschberg, D. S. (1975). A linear space algorithm for computing maximal common subsequences. *Communications of the ACM* 18(6): 341-343.

Huff, P. (2010). *Automatically Growing Language Family Trees Using the ALINE Distance*. M.A. Thesis. Brigham Young University.

Kondrak, G. (2002). *Algorithms for Language Reconstruction*. PhD Thesis. University of Toronto.

Ladefoged, P. (1995). *A Course in Phonetics*. Harcourt Brace Jovanovich. New York

Melnar, L. and C. Liu. (2006). A combined phonetic-phonological approach to estimating cross-language phoneme similarity in an ASR environment. *Proceedings of the 8th Meeting of the ACL Special Interest Group on Computational Phonology at HLT-NAACL 2006*, 1-10. New York.

Moreno, P. J, C. Joerg, J-M Van Thong, O. Glickman. (1998). A recursive algorithm for the forced alignment of very long audio segments. In *ICSLP 1998*, *Proceedings of the 5th International Conference on Spoken Language Processing.* Sydney, Australia.