# LONG AUDIO ALIGNMENT FOR AUTOMATIC SUBTITLING USING DIFFERENT PHONE-RELATEDNESS MEASURES

*Aitor Álvarez, Haritz Arzelus, Pablo Ruiz*

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

## ABSTRACT

In this work, long audio alignment systems for Spanish and English are presented in an automatic subtitling scenario. Pre-recorded contents are automatically recognized at phoneme level by language-dependent phone decoders. A dynamic-programming alignment algorithm finds matches between the automatically decoded phones and the ones in the phonetic transcription for the content's script. The accuracy of the alignment algorithm is evaluated when applying three non-binary scoring matrices based on phone confusion-pairs from each phone decoder, on phonological similarity and on human perception errors. Alignment results with the three continuous-score matrices are compared to results with a baseline binary matrix, at word and subtitle levels. The non-binary matrices achieved clearly better results. Matrix samples are given in the project's website.

***Index Terms***— Long audio alignment, phonological similarity matrices, perceptual confusion matrices, automatic subtitling.

## 1. INTRODUCTION

Due to the huge subtitling demand generated by current accessibility policies, broadcasters and subtitling companies are looking for solutions to automate subtitling.

Speech processing technologies are proving helpful in speeding up the subtitling process. A widespread approach for subtitling pre-recorded contents exploits existing text transcriptions for the content (scripts). Under this approach, automatic speech-text alignment systems recover word-level time-codes from the audio for the scripts. Although speech-text alignment is an interesting approach for automatic subtitling, aligning long audio signals is challenging, given memory demands, processing time, and the decreased reliability of the commonly employed Viterbi search algorithm when aligning long sequences.

For the present work, the successful system for long audio alignment described in [1] was taken as the basis. Their alignment method was based on Hirschberg's algorithm [2], using a binary matrix for scoring alignment operations. Our study deployed a similar algorithm, but using three types of non-binary scoring matrices, based on different phoneme-relatedness criteria. Alignment systems were developed for Spanish and English.

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

## 2. RELATED WORK

Speech-text alignment has been extensively studied. Many studies follow work by [3], where forced alignment was turned into a recursive speech recognition process, iteratively adapted to the content. Dynamic programming was used to align the hypothesis text and the reference transcript at word level. Subsequent works proposed improvements to this system: [4], [5].

A different approach, which does not require adapting the models and vocabulary, is in [1]. They developed an aligner based on Hirschberg's algorithm, a dynamic programming algorithm used in bioinformatics for genetic sequence alignment. They used a binary matrix to score alignment operations: insertions, deletions and substitutions had a cost of 1, while matches received a score of 0.

Whereas [1] used binary matrices, our study tested non-binary scoring matrices, based on phone-confusion ratios in our phone decoder, on phonological similarity, and on phone confusion in human perception. Our phonological similarity metric was based on [6], where a metric was presented that outperformed previously existing measures, applied to the task of cognate alignment. The metric was also successfully employed in spoken document retrieval [7]. Regarding phone confusion in human perception, [8] provided phone confusion results for American English, using a phoneset that is very close to our aligner's phoneset.

## 3. LONG AUDIO ALIGNMENT SYSTEM

The goal of an audio alignment system is to recover time-codes from the audio for words in the audio's script. To this end, our speech-text alignment system aligns two sequences of phonemes obtained from different sources. A language-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

dependent phone-decoder recognizes the phones and their time-codes from the audio. The decoder's output usually contains mistakes due to common recognition errors. Besides, a grapheme-to-phoneme module translates the input transcript into the reference phoneme transcription. An alignment algorithm finds phoneme matches between the phones recognized by the phone-decoder and the reference phoneme transcription. Correctly aligned phonemes will receive the time-codes obtained by the phone-decoder. Phonemes are not always correctly aligned; substitutions, deletions and insertion errors may occur. Nonetheless, the results of this study suggest that the number of matching phonemes found by our aligner is sufficient to recover enough time-codes to create near-perfectly aligned subtitles.

### 3.1. Phone recognition system

The phone recognition systems were trained using HTK, a toolkit for building hidden Markov models. The acoustic models were based on a monophone model, with three left-to-right emitting states using 32 Gaussian mixture components. The language models were bigram phoneme models. 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 recognition system was trained and tested with 20 hours of audios from three databases; Albayzín [9], Multext [10] and records of clean-speech broadcast news contents. The contents were mixed and divided into training (70%) and test (30%) sets. Texts totaling 45 million words were crawled from national newspapers to train the language model. The Phone Error Rate (PER) for the Spanish phone-decoder was 40.65%.

The English phone recognition system was trained and tested on the TIMIT database [11], which consists of 5 hours and 23 minutes of speech data. 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 PER for the English phone-decoder was 35.52%.

### 3.2. Grapheme-to-phoneme transcriptors

Two language-dependent grapheme-to-phoneme (G2P) transcriptors were developed for Spanish and English. The Spanish transcriptor was rule-based. It was inspired on the tool provided by López (*www.aucel.com/pln/*), and adapted to our phonelist. The English transcriptor was inferred from the Carnegie Mellon Pronouncing Dictionary (*svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/*) using Phonetisaurus (*code.google.com/p/phonetisaurus/*), a grapheme-to-phoneme framework driven by Weighted Finite State Transducers (WFST). The Spanish and English phonesets are available on our project's website (see *sites.google.com/site/similaritymatrices/*).

### 3.3. Algorithm for long sequences alignment

For our study, Hirschberg's algorithm was modified in two respects. (1) Given the sequences $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_m\}$ which have to be recursively divided at indexes $x_{mid}$ and $y_{mid}$ respectively, Hirschberg defined that $x_{mid}$ will always correspond to $X$'s middle index, i.e. *round(length(X)/2)*. However, for sequence $Y$, when determining the $y_{mid}$ index with Hirschberg's function, several candidate-indexes can arise. Our modification consists in always choosing as $y_{mid}$ the candidate-index that is closest to the middle of $Y$. (2) During the recursive application of the algorithm, when both sequences have a length of 1 and their phones do not match, a substitution operation is forced, in order to avoid excessive gaps. Both modifications were applied once their effectiveness was established.

In our alignment algorithm, four edit-operations are allowed: matches, substitutions, deletions and insertions. The scores for the first two operations 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. SIMILARITY MATRICES

Our scoring matrices provide the alignment algorithm with phoneme-similarity information. Depending on the scoring matrix, the alignment will be different, since scores for phoneme matches and mismatches (substitutions) differ across matrices. We developed three types of matrices, applying different phoneme-relatedness criteria. These matrices can help the alignment algorithm consider mismatches between similar phonemes as possible correct substitutions. They also prevent substitutions between very dissimilar phonemes, which are unlikely to be correct. The phone-decoder error based matrix achieves these ends by providing information about the phone decoder's phone confusion-pairs. The phonological similarity matrix is decoder-independent and estimates similarity based on common articulatory characteristics between phonemes. The perceptual matrix, also decoder-independent, reflects phone confusion-pairs in human perception. Its use is justified by the way the signal was parameterized: The frequency warping scale used for filter spacing in MFCC computation is the Mel scale, which was originally created through human perception experiments. Our project's website provides samples for the three types of matrices.

Alignment results with the three matrix-types were compared to results with a baseline binary matrix where matches had a score of 1 and mismatches had a score of 0.

### 4. 1. Phone-decoder error-based matrices

Our phone-decoder error-based matrix was computed from the output of HTK's HResults tool, when aligning the phonetic recognition and G2P transcription for sequences of ca. 25000 phonemes in Spanish and ca. 12700 phonemes in English. The matrix represents the percentage of times each phone in the phoneset was recognized correctly or misrecognized. Percentages were normalized to a 1-1000 integer range. E.g. if 8.5% of the occurrences of /θ/ were misrecognized as /f/, the matrix shows a score of 85 for phoneme pair [θ, f]. For pairs where phonemes were never mistaken for each other, we stipulated a score of –500, i.e. *½ × (0 – max({Score Range})*, preventing substitutions between members of such phoneme pairs.

### 4.2. Phonological similarity matrices

The phonological similarity scores were based on the metric described by Kondrak [6] as part of a cognate alignment system. Phonemes are described with Ladefoged's [12] multivalued features, and weighted according to their *salience*: the feature's impact for similarity. Our feature set, feature and salience values are on our project's website (see *sites.google.com/site/similaritymatrices/*).

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

where

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

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

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

**Figure 1**: Similarity function, based on Kondrak (2002)

Figure 1 shows our scoring function: $\sigma_{sub}(p, q)$ yields the similarity score for segments $p$ and $q$. $C_{sub}/100$ is the maximum possible similarity score. $C_{vwl}$ defines the relative weight of consonants and vowels. Values for $C_{sub}$ and $C_{vwl}$ are set heuristically. Function diff($p$, $q$, $f$) returns the difference between segments $p$ and $q$ for feature $f$. Feature-set $R$ is configurable. Finally, $\sigma_{skip}(p)$ returns the penalty for insertions and deletions used in the aligner (see Section 3.3).

We modified Kondrak's original function, making it more suitable for audio alignment, and for coherence with our aligner. First, the definition of *V(p)* was modified. Note that, as $C_{vwl}$, approaches 0, scores for vowel matches become closer to scores for consonant matches, increasing the weight of vowels in alignment. Kondrak mentions that prioritizing consonant matches is desirable in cognate alignment. Nonetheless, for audio alignment we obtained slightly better results by assigning the same weight to all matches. This can be achieved in the original function by setting $C_{vwl} = 0$. However, as $C_{vwl}$ approaches 0, substitutions between vowels and consonants become less clearly penalized by the matrix, which is undesirable. By adding the *or*-clause "*or p = q"* in the definition of *V(p)*, we can give equal weight to all matches, while still setting $C_{vwl} > 0$, and thus still applying an extra penalty to vowel/consonant substitutions that is not applied to vowel/vowel substitutions.

Further modifications were the following. First, adding a denominator of 100 to $\sigma_{sub}$, in order to keep Kondrak's output range, but using integer feature values and avoiding decimals to reduce memory use. Second, redefining $\sigma_{skip}$, for coherence with the way the gap penalty is calculated (see Section 3.3) when aligning with the perceptual and decoding-error based matrices.

The final modification was omitting a clause from the original function, which evaluates two-to-one phoneme alignments. These are not implemented in our audio aligner.

We defined heuristically a $C_{sub}$ value of 3500, yielding a maximum possible similarity score of 35 ($C_{sub}/100$), and a gap penalty of 9 for alignment: *ceiling(|$C_{sub}$/400|)*.

### 4.3. Perceptual similarity matrices

We created these matrices for English only. The scores were based on perceptual confusion matrices from [8]. They asked native speakers of English to identify 645 CV (ConsonantVowel) and VC syllables containing a phoneme from a 39-phoneme set (covering all of our phoneset but schwa), at signal-to-noise ratios (SNR) of 0, 8 and 16. Our scores reflect confusion percentages at SNR 16; the scoring matrix thus obtained yielded better results on our test-set than data at other SNR.

We normalized the confusion percentages for each phoneme-pair into a 1-1000 range. For phoneme-pairs where no confusion had taken place, we stipulated a score of –500, i.e. *½ × (0 – max({Score Range})*.

## 5. EVALUATION AND RESULTS

Our long audio alignment system was evaluated at word and subtitle level. The Spanish test-set (47,480 phonemes; 8,774 words and 1,249 subtitles) was composed of clean-speech audios from films. By contrast, the English test-set (21,310 phonemes; 4,732 words and 471 subtitles) consisted of non-clean speech from television audios, containing disfluencies, music, noise and overlapping speech. In addition, the English reference contained segments with imperfect transcriptions, missing subtitles for certain parts of the audio. Due to these difficulties, lower accuracy in English was expected and observed at all evaluation levels.

| *Matrix type* | *Eval Level* | *0* | *≤0.1* | *≤0.5* | *≤1.0* | *≤2.0* |
|---|---|---|---|---|---|---|
| **Binary Baseline** | WL | 14.15 | 57.82 | 72.68 | 76.20 | 79.02 |
| | SL | 10.57 | 45.24 | 73.34 | 94.96 | 100 |
| **Phone-Decoder Error-Based *(PDE)*** | WL | **23.34** | **82.28** | **94.42** | **95.99** | **97.21** |
| | SL | **18.01** | **66.85** | **87.99** | **98.96** | **100** |
| **Phonological Similarity *(PHS)*** | WL | 23.00 | 82.16 | 93.65 | 95.58 | 96.92 |
| | SL | 17.85 | 66.53 | **87.99** | 98.80 | 100 |

**Table 1**: **Spanish** word-level (**WL**) and subtitle-level (**SL**) alignment accuracy. Percentage of words and subtitles aligned within each deviation range from reference.

| *Matrix type* | *Eval Level* | *0* | *≤0.1* | *≤0.5* | *≤1.0* | *≤2.0* |
|---|---|---|---|---|---|---|
| **Binary Baseline** | WL | 0.28 | 4.81 | 19.02 | 29.67 | 43.20 |
| | SL | 0.21 | 4.03 | 36.94 | 84.08 | 100 |
| **Phone-Decoder Error-Based *(PDE)*** | WL | **2.20** | **25.73** | **52.53** | **65.35** | 76.31 |
| | SL | **0.42** | **11.46** | **48.83** | **88.32** | **100** |
| **Phonological Similarity *(PHS)*** | WL | 1.97 | 24.23 | 49.81 | 62.71 | 73.89 |
| | SL | 0.42 | 8.28 | 43.10 | 85.99 | 100 |
| **Perceptual Error-Based *(PCE)*** | WL | 1.89 | 23.76 | 50.54 | 63.93 | **76.44** |
| | SL | 0.21 | 8.92 | 47.98 | **88.32** | 100 |

**Table 2**: **English** word-level (**WL**) and subtitle-level (**SL**) alignment accuracy. Percentage of words and subtitles aligned within each deviation range from reference.

Tables 1 and 2 present the alignment accuracy at word and subtitle level for Spanish and English. We adopted the evaluation method from [3] and [1]. The cumulative percentage of correctly aligned words within several deviation ranges was recorded: Column 0 presents the percentage of perfectly aligned words, column ≤0.1 means the percentage of words correctly aligned within a maximum deviation of 0.1 sec, etc. To obtain the actual word-level time-codes for the reference files, a forced alignment was done subtitle by subtitle using the reference material, which contained subtitles manually created by professional subtitlers, as well as their time-codes. For subtitle-level evaluation, the deviation of the subtitle's first and last word compared to the reference was measured.

The most salient conclusion supported by the results is that using matrices based on phone-decoder errors (PDE), phonological similarity (PHS) or perceptual errors (PCE) significantly improved alignment accuracy compared to using the binary matrices. The improvements are noticeable at all deviation ranges.

For both languages, the best alignment accuracy was obtained using the PDE matrix. This was expectable, since the matrix is based on phone confusion-pairs from the phone-decoder used by the aligner.

In Spanish, with the PDE matrix, accuracy gains of 21 and 14 percentage points (ptp) were obtained at subtitle-level compared to the binary matrix, at 0.1 and 0.5 second deviations respectively. Considering that 0.1 and 0.5 sec. are acceptable deviations for subtitling applications, these gains represent a positive impact at an actual application level. Improvements were even higher in word-alignment accuracy: 37 ptp and 21 ptp at the same deviation ranges.

For English, alignment accuracies are lower, given difficulties posed by the test-set. However, the improvements with the PDE matrix compared to the binary matrix are also clear: 7 ptp and 12 ptp at subtitle level with 0.1 and 0.5 second deviations respectively, while at word level accuracies reached improvements of 21 ptp and 33 ptp for the same deviation ranges.

Regarding the PHS and PCE matrices, their alignment accuracy was close to the accuracy obtained with the PDE matrix. This finding suggests that the performance of decoder-independent matrices can get close to the performance of decoder-dependent matrices. Also note that, even if both PHS and PCE matrices obtained similar results, there was a small trend for the PCE matrix to be more accurate. The nature of the PCE matrix, more closely-related to the signal parametrization than the PHS matrix, could explain these minimal accuracy differences.

## 6. CONCLUSIONS AND FUTURE WORK

Several scoring matrices, based on different phoneme-relatedness criteria, were tested for aligning long audios with Hirschberg algorithm, and found to improve alignment accuracy compared to a binary matrix. As expectable, the matrix based on phone-decoder errors achieved the most accurate alignment results, while the matrices based on phonological similarity and perception errors also obtained clear improvements in alignment accuracy. Improvements were observed at word and subtitle level for Spanish and English, even if the English test-set posed serious difficulties. Thus, the effectiveness of the matrices under adverse conditions was also established.

Accuracy does not only depend on the alignment algorithm and the scoring matrices, but also on the performance of the phone decoders. Improving their robustness, by using a larger training-set or adapting models to the content or domain, will increase alignment accuracy.

Other future work would be extending the system to more languages. For this study, a matrix based on perceptual errors was created for English, but not for Spanish.

Regarding Hirschberg's algorithm, it sometimes offers more than one possible optimal alignment. In this study, we forced the algorithm to compute just one solution. Considering the different possible solutions, and defining criteria to choose among them could be an interesting study.

## 7. REFERENCES


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

[2] D. S. Hirschberg, "A linear space algorithm for computing maximal common subsequences.," *Communications of the ACM,* vol. 18, no. 6, pp. 341-343, 1975.

[3] P. J. Moreno, C. Joerg, J-M Van Thong and O. Glickman, "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.

[4] P. J. Moreno and C. Alberti, "A factor automaton approach for the forced alignment of long speech recordings," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Taipei, Taiwan, 2009.

[5] A. Katsamanis, M. P. Black, P. G. Georgiou, L. Goldstein and S. Narayannan, "SailAlign: Robust long speech-text alignment," in *Proceedings of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[6] G. Kondrak, Algorithms for Language Reconstruction, PhD Thesis. University of Toronto, 2002.

[7] P. Comas, Factoid Question Answering for Spoken Documents, PhD Thesis. Universitat Politècnica de Catalunya, 2012.

[8] A. Cutler, A. Weber, R. Smits and N. Cooper, "Patterns of English phoneme confusions by native and non-native listeners," *Journal of the Acoustical Society of America,* vol. 116, no. 6, pp. 3668-3678, 2004.

[9] J. E. Díaz, A. Peinado, A. Rubio, E, Segarra, N. Prieto and F. Casacubierta, "Albayzín: a task-oriented Spanish speech corpus," in *Proceedings of the First International Conference on Language Resources and Evaluation*, *LREC,* Granada, Spain, 1998.

[10] E. Campione and J. Véronis, "A multilingual prosodic database," in *Proceedings of the 5th International Conference on Spoken Language Processing*, *ICSLP,* Sydney, Australia, 1998.

[11] J. S. L. Garafolo, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, Philadelphia, 1993.

[12] P. Ladefoged, A Course in Phonetics, New York: Harcourt Brace Jovanovich, 1995.