vicomtech
visual interaction & communication technologies

Phoneme Similarity Matrices to Improve Long Audio Alignment for Automatic Subtitling

Pablo Ruiz, Aitor Álvarez, Haritz Arzelus
{pruiz,aalvarez,harzelus}@vicomtech.org

Vicomtech-IK4, Donostia/San Sebastián, Spain
www.vicomtech.org

***LREC,* Reykjavik 28 May 2014**

Introduction

THE NEED

Huge subtitling demand by broadcasters, for accessibility compliance: Automatic subtitling is an attractive option.

TALK OUTLINE

- Alignment system for automatic subtitling
  - Phone decoder
  - Grapheme-to-Phoneme transcriber
  - Alignment algorithm
- Alignment Results applied to Subtitling

Alignment approach

- Long Audio Alignment: Aligns the audio signal with a human transcript for the audio.
- System by Bordel et al. (2012):
  - Alignment with Hirschberg's algorithm.
  - Simple system, but accuracy comparable to more complex approaches (cf. Moreno et al. 1999)
  - Uses a BINARY scoring matrix to evaluate alignment operations.
- Our system: Hirschberg's algorithm, using NON-BINARY scoring matrices, improving *vs.* binary

Speech-Text Alignment System

Phone Decoder

G2P

Phones

Time-codes (TC)

Hirschberg's Algorithm

Similarity Matrices

Aligned Phones + TC

Aligned Words and Subtitles

Phone Decoder

Acoustic Model (AM): HTK – Monophone (18 MFCC + *Δ* + *ΔΔ*)

Bigram Phoneme Language Model (LM)

Languages	Train	Test	PER
Spanish	~15H (*LM 45 M*)	~5H	40.65%
English	~4H (*LM 369M*)	~1H	35.52%

Sources for Corpora

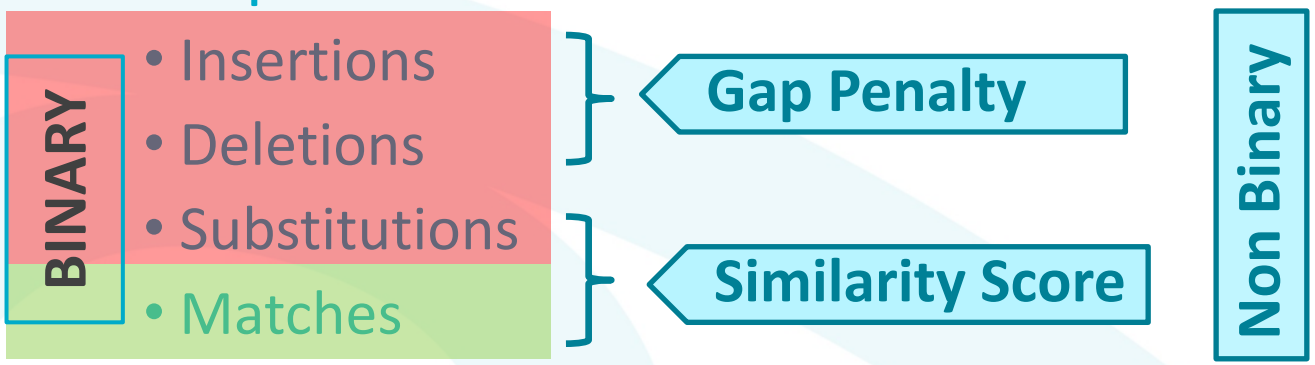
- Spanish: Albayzin, Multext, SAVAS (AM), *newspapers (LM)*
- English: TIMIT (AM), *newspapers (LM)*

Grapheme-to-Phoneme (G2P) Transcriptors

- Spanish: rule-based
- English: inferred from CMU Dictionary using Phonetisaurus (WFST)
- Phone-sets and details: https://sites.google.com/site/similaritymatrices

Alignment Algorithm

- Hirschberg's algorithm for sequence alignment
  - –Dynamic programming, optimization to Needleman-Wunsch, applicable to longer sequences.
  - –Four operations



  - –Each operation is evaluated with a scoring function
    - Binary vs. non-binary
    - Goal: Promotes aligning equal or similar, easily confusable, phonemes. Prevents aligning dissimilar phonemes.

Speech-Text Alignment (recap)

Phone Decoder

G2P

Phones

Time-codes (TC)

Hirschberg's Algorithm

Similarity Matrices

Aligned Phones + TC

Aligned Words and Subtitles

Similarity Function: Input Features and Values

- Phonological similarity: Multivalued features weighted by salience (i.e. impact in similarity)

IPA	Place[1]		Manner[1]		V	Syl	Voi	Nas	Lat	Asp	High[1]		Back[1]		Ro[1]	Lo[1]
æ	palatal	70	low vowel	0	1	100	100	0			low	0	front	100	0	0
i:	palatal	70	high vowel	40	1	100	100	0			high	100	front	100	0	100
n	alveolar	85	stop	100	0	0	100	100	0	0						
p	bilabial	100	stop	100	0	0	0	0	0	100						
ɹ	alveolar	85	approximant	60	0	0	100	0	0	0						
s	alveolar	85	fricative	80	0	0	0	0	0	0						
aj	palatal	70	low vowel+ high vowel	16	1	100	100	0			low+ high	40	central+ front	70	0	100

Feature Values (English)

Feature Salience

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

Similarity Function: Definition (*cf.* Kondrak 2002)

σ_{sub}: score for substituting phoneme ***p*** with ***q***

1 $$\sigma_{sub}(p,q) = (C_{sub} - \delta(p,q) - V(p) - V(q)) / 100$$

2 where if $p = q$, $V(p) - V(q) = 0$

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

4 $$\delta(p,q) = \sum_{f \in R} |\text{diff}(p,q,f)| \times \text{salience}(f)$$

σ_{skip}: cost of skipping

5 $$\sigma_{skip}(p) = |C_{skip} / 100|$$

$C_{sub} = 3500$
$C_{vwl} = 1000$
$C_{skip} = 1000$

	a	b	d	ɛ	f	ʌ	l	n	o	p	r	r	θ	t	ʧ
a	35	-46	-42	10	-44	-24	-30	-50	4	-56	-30	-20	-42	-52	-49
b	-46	35	31	-32	13	-7	-1	19	-36	25	-1	9	11	21	-2
d	-42	31	35	-28	13	-3	3	23	-32	21	3	13	15	25	2
ɛ	10	-32	-28	35	-30	-10	-16	-36	4	-42	-16	-6	-28	-38	-35
f	-44	13	13	-30	35	-5	1	1	-34	23	1	11	33	23	10
ʌ	-24	-7	-3	-10	-5	35	29	-11	-14	-17	9	19	-3	-13	-10
l	-30	-1	3	-16	1	29	35	-5	-20	-11	15	25	3	-7	-16
n	-50	19	23	-36	1	-11	-5	35	-40	9	-5	5	3	13	-6
o	4	-36	-32	4	-34	-14	-20	-40	35	-46	-20	-10	-32	-42	-39
p	-56	25	21	-42	23	-17	-11	9	-46	35	-11	-1	21	31	8
r	-30	-1	3	-16	1	9	15	-5	-20	-11	35	25	3	-7	-16
r	-20	9	13	-6	11	19	25	5	-10	-1	25	35	13	3	-6
θ	-42	11	15	-28	33	-3	3	3	-32	21	3	13	35	25	12
t	-52	21	25	-38	23	-13	-7	13	-42	31	-7	3	25	35	12
ʧ	-49	-2	2	-35	10	-10	-16	-6	-39	8	-16	-6	12	12	35

	a	b	d	ɛ	f	ʌ	l	n	o	p	r	r	θ	t	ʧ
a	35	-46	-42	10	-44	-24	-30	-50	4	-56	-30	-20	-42	-52	-49
b	-46	35	31	-32	13	-7	-1	19	-36	25	-1	9	11	21	-2
d	-42	31	35	-28	13	-3	3	23	-32	21	3	13	15	25	2
ɛ	10	-32	-28	35	-30	-10	-16	-36	4	-42	-16	-6	-28	-38	-35
f	-44	13	13	-30	35	-5	1	1	-34	23	1	11	33	23	10
ʌ	-24	-7	-3	-10	-5	35	29	-11	-14	-17	9	19	-3	-13	-10
l	-30	-1	3	-16	1	29	35	-5	-20	-11	15	25	3	-7	-16
n	-50	19	23	-36	1	-11	-5	35	-40	9	-5	5	3	13	-6
o	4	-36	-32	4	-34	-14	-20	-40	35	-46	-20	-10	-32	-42	-39
p	-56	25	21	-42	23	-17	-11	9	-46	35	-11	-1	21	31	8
r	-30	-1	3	-16	1	9	15	-5	-20	-11	35	25	3	-7	-16
r	-20	9	13	-6	11	19	25	5	-10	-1	25	35	13	3	-6
θ	-42	11	15	-28	33	-3	3	3	-32	21	3	13	35	25	12
t	-52	21	25	-38	23	-13	-7	13	-42	31	-7	3	25	35	12
ʧ	-49	-2	2	-35	10	-10	-16	-6	-39	8	-16	-6	12	12	35

	a	b	d	ɛ	f	ʌ	l	n	o	p	r	ɾ	θ	t	ʧ
a	35	-46	-42	10	-44	-24	-30	-50	4	-56	-30	-20	-42	-52	-49
b	-46	35	31	-32	13	-7	-1	19	-36	25	-1	9	11	21	-2
d	-42	31	35	-28	13	-3	3	23	-32	21	3	13	15	25	2
ɛ	10	-32	-28	35	-30	-10	-16	-36	4	-42	-16	-6	-28	-38	-35
f	-44	13	13	-30	35	-5	1	1	-34	23	1	11	33	23	10
ʌ	-24	-7	-3	-10	-5	35	29	-11	-14	-17	9	19	-3	-13	-10
l	-30	-1	3	-16	1	29	35	-5	-20	-11	15	25	3	-7	-16
n	-50	19	23	-36	1	-11	-5	35	-40	9	-5	5	3	13	-6
o	4	-36	-32	4	-34	-14	-20	-40	35	-46	-20	-10	-32	-42	-39
p	-56	25	21	-42	23	-17	-11	9	-46	35	-11	-1	21	31	8
r	-30	-1	3	-16	1	9	15	-5	-20	-11	35	25	3	-7	-16
ɾ	-20	9	13	-6	11	19	25	5	-10	-1	25	35	13	3	-6
θ	-42	11	15	-28	33	-3	3	3	-32	21	3	13	35	25	12
t	-52	21	25	-38	23	-13	-7	13	-42	31	-7	3	25	35	12
ʧ	-49	-2	2	-35	10	-10	-16	-6	-39	8	-16	-6	12	12	35

Evaluation: Method

- Alignment accuracy at word level and subtitle level compared to manual subtitles created by professional subtitlers

Alignment Test Corpora		
	SPANISH	**ENGLISH**
Words	8,774	4,732
Subtitles	1,249	471
Content	Clean speech	Noisy speech Some reference subtitles missing

Evaluation: Results at Word Level

Percentage of words aligned within a given deviation range from reference

Spanish – Word Level					
seconds	**0**	**≤0.1**	**≤0.5**	**≤1.0**	**≤2.0**
BinaryBaseline	14.17	57.71	72.65	76.21	79.02
PhonologicalSimilarity	**+8**[1]	**+24**	**+20**	**+19**	**+18**

English – Word Level					
seconds	**0**	**≤0.1**	**≤0.5**	**≤1.0**	**≤2.0**
BinaryBaseline	0.28	4.81	19.04	29.69	43.24
PhonologicalSimilarity	**+1.6**	**+19**	**+30**	**+32**	**+29**

[1]Gain in percentage *points*

Evaluation: Results at Subtitle Level

Percentage of subtitles aligned within a given deviation range from reference

Spanish – Subtitle Level					
seconds	**0**	**≤0.1**	**≤0.5**	**≤1.0**	**≤2.0**
BinaryBaseline	10.57	45.08	73.26	95.12	100
PhonologicalSimilarity	**+7[1]**	**+21**	**+14**	**+3.5**	**=**

English – Subtitle Level					
seconds	**0**	**≤0.1**	**≤0.5**	**≤1.0**	**≤2.0**
BinaryBaseline	0.21	4.25	37.15	84.29	100
PhonologicalSimilarity	**+0.2**	**+4**	**+9**	**+2.5**	**=**

[1]Gain in percentage *points*

Conclusions and Further Work

- Continuous scoring matrices based on phonological similarity improve alignment compared to a binary matrix.
- Further work: Other similarity criteria can also improve results.

ICASSP (May 2014), *Text, Speech and Dialogue* (Sept 2014)

English – Subtitle Level					
seconds	0	≤0.1	≤0.5	≤1.0	≤2.0
Binary	0.21	4.03	36.94	84.08	100
PhonologicalSim	**+0.2**	**+4**	**+6**	**+2**	**=**
DecoderErrorBased	**+0.2**	**+7**	**+12**	**+4**	**=**

Thank you

https://sites.google.com/site/similaritymatrices