Searching Relevant Syllable Context by Clustering for Alignment in Modern Greek Speech

R. Rispoli, C. Kotropoulos and I. Pitas

Dept. of Informatics, Aristotle University of Thessaloniki,
Box 451 Thessaloniki, GR-54124 GREECE
{*rispoli,costas,pitas*}*@zeus.csd.auth.gr*

Abstract

We present in this paper some results on the optimal design of syllable databases for Modern Greek. We show that speaker gender or stress are not crucial criteria for voiced syllables, whereas distance to silence has to be taken into account. Such syllable databases are exploited in an alignment system based on Dynamic Time Warping for Modern Greek speech. The overall architecture of the alignment system is also presented.

1. Introduction

It is obvious that a thorough understanding of what speech is will not be achieved without having studied a sufficient diversity of languages, and not only the strongest ones (English, Spanish, etc.) but also weak languages such as Greek. Alignment is a necessary preprocessing step in speech analysis. It consists in a set of processes returning the time boundaries of each phoneme in an utterance, when a sound file of the utterance and a phonetic transcription are given. For the time being there is no alignment system for modern Greek. An automatic alignment system would be useful to study prosodic patterns in spontaneous speech as well as in multimodal research such as talking heads synthesis. In order to realize such a system using a new combination of techniques based on samples, we have to find the best possible design for syllable databases. This paper presents some results on voiced syllable samples. We first present the outline of our alignment system and explain why focus on voiced segments is made. Then we describe our experimental procedure to find similar samples, based on automatic clustering, and finally we expose our results. It is shown that redundancies in databases may be avoided by choosing samples that have a wide variety of distance to silence.

This research is funded by the European RTN on Multimodal Human-Computer Interaction (MUHCI), HPRN-CT-2000-00111.

2. Towards a General alignment system for modern Greek

To some extent, alignment is easier than automatic speech recognition, but if accurate results are needed and if spontaneous speech is dealt with, a good design is far from being straightforward.

2.1. Outline of the system

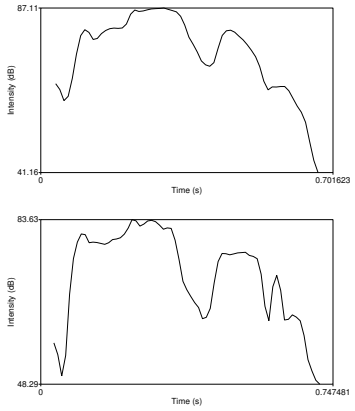

Fig. 1. Intensity profiles for different samples of “pull-over”.

Alignment systems are often designed using Hidden Markov Models and the results are generally quite good, for example [6] for German language. Nevertheless, to design such a system for modern Greek, we resort to a combination of Dynamic Time Warping (DTW) on intensity profiles with a majority voting. Main reasons for this choice are:

- Intensity profiles are often similar for a same utterance, even if the utterance is said in different contexts. For instance in Figure 1, the word “pull-over” is pronounced by a woman at the end of a sentence in the

first case, and by a man at the beginning of a sentence in the second;

- Phoneme databases underlying majority voting techniques used here could be used for speech synthesis with tools like M'BROLA in the close future.

The general architecture of our alignment system is as shown in Figure 2.

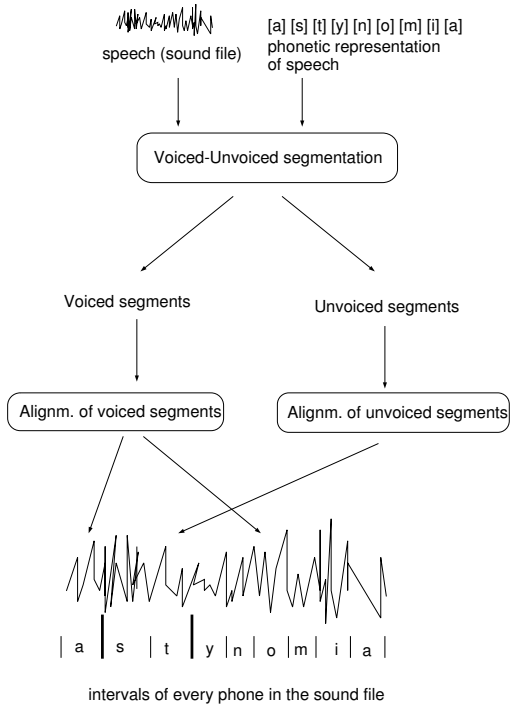

Fig. 2. Outline of the alignment system.

We first make a voiced-unvoiced separation of the signal. This separation reduces the complexity of the analysis and at the same time improves precision, as unvoiced speech parts are detected easily. Unvoiced and voiced segments are then processed separately. Voiced segments are processed by comparing their intensity shape together with first-order derivative to a template retrieved from a database of voiced phonemes. Alignment is realized using the standard asymmetric DTW procedure as phoneme segmentation is known for templates. Details are given in the next section.

As far as unvoiced phonemes are concerned, multiple combinations have been reported for Modern Greek by [2]: 16 two-consonant clusters e.g. [ps] in "ps"ari" (fish), [fk], [sf], [st] etc. and only 2 three-consonant clusters, [kst] like in ""ekstasi" (ecstasy) and [fst]. The method for aligning these segments is not decided yet but the problem seems to be an easy one because for voiced segments the intensity profiles are very characteristic for these combinations and are shown to be less context-sensitive. Therefore, the crucial point is the alignment of voiced segments, which is detailed subsequently.

2.2. Alignment of voiced segments

To align voiced segment, we use a classical warping method, i.e. we compare the unknown input with a known template. Considering that the results are improved using not only a single but multiple templates and then combining the results, the need arises to define a good template set. Automatic clustering helps to define such a template set.

2.2.1. Dynamic Time Warping

DTW is a family of algorithms returning a mapping and a distance between two utterances represented by vectors : a template $T(n)$, i.e. a known reference for the utterance and an input $I(n)$ representing the same contents, the latter being studied. The result is given as a distance between the two representations, together with two index sequences $K(t)$ and $L(t)$, meaning that sample $T(K(t))$ corresponds to $I(L(t))$ for any t. In our case, a DTW algorithm needs only the distance between $T(j)$ and $I(k)$ for any j and k, and some constraints on result sequences, such as strict monotonicity.

The use of DTW for alignment is obvious: knowing that $T(n_0)$ is a transition sample for the template T (for instance it could be the transition between phoneme [m] and phoneme [a] in a realization of [ma]), it is natural to think that $I(L(t_0))$, with $K(t_0) = n_0$, will represent the same transition in the input I. This is represented in Figure 3. Details for DTW can be found in [3], [5].

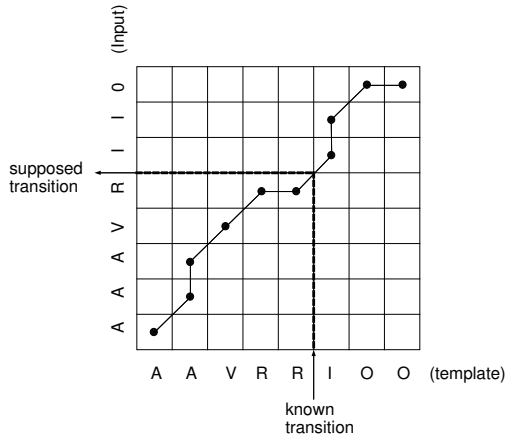

Fig. 3. DTW for alignment.

The error (in samples) is defined as the difference of what is returned by DTW procedure, i.e. $L(t_0)$ (with $K(t_0) = n_0$) and what is supposed to be the real transition, let us call

it m_0. We can reduce the error if instead of using a single template for aligning an unknown input we employ more templates.

2.2.2. Majority voting and syllable database

We use a template set and keep as result the $L(t_0)$ which is most often returned by multiple DTW procedures. In the case of a tie, the closest result to the mean can be chosen. Now the more important question is to build an optimal template database, for every possible voiced segment. In particular, we have to avoid redundancy, i.e. having many similar templates, with respect to the distance used in the database.

2.2.3. Clustering

In order to discover the samples that are similar, we used an automatic clustering method together with the distances provided by the DTW procedure. More precisely, considering that we have n templates to cluster, the use of DTW first supplies the distances between these templates taken two by two (i.e. a distance matrix $n \times n$). From these distances, a hierarchical cluster tree is created and finally we can determine where to divide the tree into clusters.

3. Experimental Procedure

A book for modern Greek learning was used that contains a CD with the spoken version of the utterances [4]. The speech speed is normal according to native speakers, and sometimes quite fast, except two or three sparse examples which are slower. All CD files were converted into .wav files (around 70 minutes are available) with 16000 Hz sampling frequency and the book was thoroughly analyzed by OCR to optimize work. We chose some samples for the same voiced syllable. Table 1 depicts the samples used for syllable [vε]. We then converted the signal into a sequential two-dimensional vector (intensity and first-order derivative). Intensity was calculated with a classical windowed Fourier Transform using a window of 15 ms duration that was shifted by 5 ms. Once every sample was represented by this way, we used an asymmetrical DTW procedure (aDTW) as in [1]. Let us note that in Table 1, we indicate some information such as gender, whether the syllable is stressed or not, and the position of the syllable in the utterance that will be discussed in next section. In Table 1, *position* is defined as follows: $pos(x) = \frac{s_x}{n+1}$ where n is the total number of the utterance between two silences and s_x is the rank of syllable x in the utterances. For instance the position of syllable [ma] in utterance [mi$\int\varepsilon$lmabεl] is $\frac{3}{4+1} = 0.6$.

For a set containing n samples, aDTW applied to every possible pair of samples provides a square matrix $n \times n$ where the diagonal coefficients are equal to 0. Then, a distance between every pair can be computed, summing the symmetrical coefficients in the matrix. As asymmetrical DTW is used, $D_{ij} \neq D_{ji}$. With such information, built-in clustering procedures can be used to generate interesting groups. In order to find clusters in the hierarchical tree, we used 0.5 as inconsistency coefficients in Matlab (every link is given an inconsistency value between 0 and 1; the lower the value, the more natural the link is. See procedures linkage, cophenet and cophenet in [8]).

4. Results

We give here the results concerning a set of samples for syllable [vε], which is represented by the Greek graphemes $\beta\varepsilon$, $\beta\alpha\iota$ and some combinations containing $\varepsilon\upsilon$. Examples are shown in Table 1. Clusters derived by the described procedure are reported in Table 2. As we used a low inconsistency coefficient, 0.5 (i.e. a restrictive one) it happens that the resultant groups contain only 2 elements. This is of course not always the case.

If we consider in Table 1 the proportion of clusters whose elements correspond to the same speaker *gender*, we obtain a ratio of $\frac{6}{9} = 0.66$. This is close to the probability P_1 of picking 2 same gender elements when randomly picked, which is 0.54. Indeed, considering that in Table 1 we have 18 sentences pronounced by female speakers and 9 by male ones, we have $P_1 = \frac{\binom{18}{2}+\binom{9}{2}}{\binom{18+9}{2}} = 0.54$. For *stress* criterion, the result, $\frac{4}{9} = 0.44$, is even closer to randomness that corresponds to $P_2 = 0.48$, since we have 14 stressed samples and 13 not stressed. This leads us to think that gender and stress are not necessarily crucial criteria for grouping, for the signal representation and the distance used.

On the other hand, if we define the distance of two samples a and b in a classical way $d(a, b) = |pos_a - pos_b|$ where pos_x is the position of syllable x in the utterance, we notice in Table 2, that distances between grouped samples are always (except for row 3) less than $P_3 = 0.33$ which is the mean distance of two points randomly chosen between 0 and 1. Actually, the mean of column *distance* in Table 2 is 0.17, which is significantly less than P_3.

Other experiments have been done for [vi] and [va], amongst others, and results are comparable.

5. Conclusions

We demonstrated in this study that in order to build a non redundant syllable database for an alignment system in modern Greek, we should care more about the distance to silence than the stress or speaker gender. Although we have worked only on a simple voiced syllable database for the

Id	speech segment (transcribed greek [SAMPA])	english translation	gender	stress	position
1	ba, ”kati sim”*ve*ni	Bah, something happens	f	yes	0.66
2	kate”*ve*nis sti ”stasi	you get off at the station	f	yes	0.43
3	mi”lo ce katala”*ve*no elini”ka	I speak and understand Greek	f	yes	0.54
4	”spania pao se ta”*ve*rnes	I go barely to taverns	f	yes	0.75
5	pi”jjenis se ta”*ve*rnes	are you going to taverns	m	yes	0.75
6	”*ve*vea	of course	f	yes	0.25
7	ta kondo”manika ”*ve*vea	short sleeves of course	f	yes	0.70
8	Ta ”fame se mia psarota”*ve*rna	we’ll have lunch in a fish-tavern	f	yes	0.50
9	pu jje”niTike sti ”*ve*ria	who was born in Veria	m	yes	0.35
10	”oCi ”*ve*vea	of course not	m	yes	0.83
11	ne ”*ve*vea	yes, of course	m	yes	0.40
12	ne ”*ve*vea	yes, of course	f	yes	0.40
13	”*ve*vea, e”si;	of course, what about you?	f	yes	0.17
14	asti”evese ”*ve*vea	you must be joking	m	yes	0.63
15	Ta ”paro tiz ”vla*ve*s tu o”te	I will call the telecom hot line	f	no	0.65
16	a”komi ce pra”tiria *ve*n”zinis	and even fuel stations	f	no	0.77
17	Min kse”xasis ce ka”nena pu”lo*ve*r	and don’t forget to take a sweater	f	no	0.94
18	pu”lo*ve*r Den ”perno	I’m not taking a sweater	m	no	0.43
19	e”Go ”leo ”ena ”malino pu”lo*ve*r	I propose a woollen sweater	m	no	0.92
20	ka”tala*ve*s	do you understand?	f	no	0.80
21	*ve*”veos ki”ria Anasta”siu	Of course Mrs Anastasiou	f	no	0.08
22	*ve*”veos bo”ri	of course it can	f	no	0.17
23	xo”re*ve*te;	do you dance?	m	no	0.60
24	i ki”dia tu Sera”feim Dimosi”e*ve*te	Seraphin’s funeral is announced	f	no	0.52
25	asti”e*ve*se ”vevea	you must be joking	m	no	0.38
26	ti oni”re*ve*se;	what is your dream	f	no	0.71
27	pan”dre*ve*te o kir i”lias	Mr Ilias got married	f	no	0.30

Table 1. Material used for [vε] (studied syllable in *italic*)

Natural clusters (Ids from Table 1)	Same gender	Same stress	distance
2,3	yes	yes	0.11
4,20	yes	no	0.05
5,12	no	yes	0.35
7,17	yes	no	0.24
8,27	yes	no	0.20
18,24	no	yes	0.09
11,26	no	no	0.31
23,25	yes	yes	0.22
22,13	yes	no	0.00

Table 2. Clustering results

time being, the results seem to be general enough to be extended. We are verifying this hypothesis for more complex voiced structures, such as syllables containing diphthongs (e.g. [dia]) and/or many consonants ([vrε]). We are also presently studying the critical number of samples for a given syllable. Next step will be the design of a merging process for the stored templates. An overall alignment system will be then close to completion.

6. References

[1] S. Wrigley (1999), Speech Recognition by Dynamic Time Warping, University of Sheffield (http://www.dcs.shef.ac.uk/ stu/com326/).

[2] D. Holton, P. Mackridge and I. Philippaki-Warburton (1997), Greek: A comprehensive Grammar of the Modern Language. Routledge.

[3] J. Picone, K. Goudie-Marshall, G. R. Doddington and W. Fisher, “Automatic Text Alignment for Speech System Evaluation”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, August 1986.

[4] E. Demiri and R. Kamarianou (2002), Nea Ellinika gia metanastes, Alfa epipedo (Modern Greek for immigrants, A-level), Metaixmio.

[5] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, February 1978.

[6] S. Rapp, Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models / An Aligner for German. In: Workshop ”Integration of Language and Speech in Academia and Industry”, Moskow, November 1995. ELSNET goes east and IMACS.

[7] P. Boersma and D. Weenink (2003), Praat: doing phonetics by computer. (http://www.fon.hum.uva.nl/praat/Praat)

[8] Matlab Statistics Toolbox Reference Manual. (http://www.mathworks.com/)