Semi-Automatic Aligning of Swedish Forensic Phonetic Phone Speech in Praat using Viterbi Recognition and HMM

Jonas Lindh, Department of Linguistics,
Göteborg University, Sweden.

jonas@ling.gu.se

Abstract

Automatically aligning sound and text on phone, syllable, word and phrase level is a valuable tool. Handling speech databases of various kinds and developing speech technology tools most often demands some kind of aligning. Using the free software Praat a plugin framework for automatic aligning, Easyalign, has been developed. With a Swedish grapheme to phone converter, a Swedish trained hidden markov model and the viterbi function HVite from the toolkit HTK, automatic aligning of an authentic forensic phonetic recording and corresponding orthographic transcription was produced. The result, to some extent successful, and conclusion invites to more research and developments for the future.

Background and Introduction

To automatically align text and sound is enormously valuable when one is handling large speech databases, either for research or for developing speech technology tools such as automatic speech recognition or text to speech systems. It is also a very useful tool in forensic speaker identification as one often receives a tapped recording together with an orthographic transcription. This orthographic transcription can then be used together with the sound file to get a crude overview of where in the recording certain events occur according to the orthographic transcription. Even if the aligning sometimes is very crude, it certainly facilitates the tedious manual work of labeling and transcribing. To be able to perform automatic aligning common speech recognition techniques are applied at various levels. In this case, a framework for

doing automatic aligning, called Easyalign, was developed for the free software Praat (Goldman, 2007). Praat is distributed as an open source software under a GPL license. On top of the source code a built in scripting language can execute commands, make calculations and communicate with other programs in different manners (Boersma & Weenink, 2007). To be able to implement a new language for automatic aligning within the framework there is a need for some kind of grapheme to phone converter and a trained Hidden Markov Model that can be used by the viterbi recognition program HVite from the HTK toolkit (Young et al., 2006).

Automatic Aligning of Speech

Aligning recorded speech automatically is a technique that borrows heavily from automatic speech recognition (ASR). Successful attempts have been made using hidden markov models (Brugnara et al., 1993) and dynamic time warping (Malfrére et al., 1998 and 2000), both well-known techniques in ASR. When it comes to dynamic time warping, the signal is compared and aligned with a reference from for example a text to speech system. Using a hidden markov model (hmm) recognition system, forced alignment can be used together with phoneme models and the Viterbi algorithm. (Sjölander, 2001 and 2003). The output of the forced alignment can then be used to create other tiers on other phonological levels. The result can then be displayed together with the sound in software that can read a certain labeling format, such as Praat.

Praat and EasyAlign

Praat is free software for analysis, synthesis and manipulation of speech. It is also possible to create pictures and graphs to illustrate for example an analysis (Boersma & Weenink, 2007). One of the many advantages of this open source software is the ability of scripting. By scripting one can implement formulas and easily use the different functions built into Praat. It is also possible to communicate with other scripts or just execute system commands applying other functions, which is used in this case. Since Praat is built for phonetic analysis, it also contains built-in functions

to display and manually label sound files. Textgrids with desired amount of layers can easily be produced, where different kinds of information can be displayed at the same time depending on the corpus one is working with or creating. These functions are very well suited for displaying and editing the results of automatic alignment. For this purpose EasyAlign was created. EasyAlign is a framework for building the possibilities for automatic aligning of speech with many different languages (Goldman, 2007). EasyAlign is based on several different Praat scripts, the viterbi function program HVite from the HTK toolkit, as a binary executable (Young et al., 2006), a grapheme to phone (G2P) converter and a trained HMM for each language. As different phonetic transcription alphabets are applied for different purposes there is a mapping table applied for each phonetic conversion and if there is a difference between the G2P output and the transcription used in the HMM. First, the orthographic text is aligned at the phrase level with the sound file based on a pausing threshold given (here 90 ms) and the punctuation marks in the text in the same way a TTS system applies pausing for the synthesis. The second step is to convert the orthographic string to a phonetic one by executing the G2P depending on the language involved. All steps can the directly be displayed in a separate window for manual correction or inspection. The last step is to apply the viterbi function for each phrase together with the phonetic string and the HMM and then use the phone level output to also give a syllable and word level alignment.

Method

As described above both Praat and EasyAlign are very well suited for implementation of a Swedish automatic aligner. However, first some kind of G2P had to be built that could be easily called from within Praat together with a trained HMM together with some corrections and modifications of scripts to suit both audio quality, Swedish orthography and transcription.

Phonetic Transcription

In EasyAlign, SAMPA is used for the so far implemented languages French and Hebrew. In order not to deviate too much from the previously developed implementations a Swedish SAMPA (http://www.phon.ucl.ac.uk/home/sampa/index.html) transcription was used with a few modifications to avoid misinterpretations between the orthographic and phonetic strings and the use of uncommon symbols. Below is the phonetic transcription used for this project (SAMPA EA-Jonas) in a table together with the common SAMPA notation for Swedish.

SAMPA (EA-Jonas)	**SAMPA (Swedish)**	**Word**	**Transcription**	**Comment**
Consonants				
There are six plosives:				
p	p	pil	pi:l	
b	b	bil	bi:l	
t	t	tal	tA:l	
d	d	dal	dA:l	
k	k	kal	kA:l	
g	g	gås	go:s	
There are six fricatives:				
f	f	fil	fi:l	
v	v	vår	vo:r	
s	s	sil	si:l	
S	S	sjuk	S}:k	(front and back allophones#)
h	h	hal	hA:l	
C	C	tjock	COk	(not syllable-final)
There are six sonorant consonants (nasals, liquids and semivowels):				
m	m	mil	mi:l	
n	n	nål	no:l	
N	N	ring	rIN	(not syllable-initial)
r	r	ris	ri:s	
l	l	lös	l2:s	
j	j	jag	jA:g	
Vowels				
There are nine long and nine short vowels.				
Long vowels (followed by short consonant):				
i:	i:	vit	vi:t	
e:	e:	vet	ve:t	
E:	E:	säl	sE:l	

y:	y:	syl	sy:l	
uh:	}:	hus	h}:s	
ox:	2:	föl	f2:l	
u:	u:	sol	su:l	
o:	o:	hål	ho:l	
A:	A:	hal	hA:l	
Short vowels (followed by long consonant):				
I	I	vitt	vIt	
e	e	vett	vet	Merging categories
e	E	rätt	rEt	Merging categories
Y	Y	bytt	bYt	
u0	u0	buss	bu0s	
ox	2	föll	f2l	
U	U	bott	bUt	
O	O	håll	hOl	
a	a	hall	hal	
There are also two pre-r allophones (long and short) of /E/ and /2/ (see below).				
The following important allophonic variants occur in Swedish which require separate symbolic representation:				
ae:	{:	här	h{:r	pre-r allophone of E:
oe:	9:	för	f9:r	" 2:
ae	{	herr	h{r	" E
oe	9	förr	f9r	" 2
eh	@	pojken	pOjk@n	schwa vowel allophone
T	rt	hjort	jUrt	retroflex consonant, not initial*
D	rd	bord	bu:rd	"
2N	rn	barn	bA:rn	"
2S	rs	fors	fOrs	"
L	rl	karl	kA:rl	"

Rather crude rules for G2P were then created using regular expressions in a Perl program. The Perl program was then converted into a binary executable for windows and implemented into the EasyAlign framework.

Hidden Markov Models Trained on a Swedish Corpus

The Swedish SpeechDat project was a part of a larger project to create databases of recorded telephone speech to be able to train and develop applications for speech recognition and verification (Elenius et al., 1997). The database outcome was

telephone recordings for approximately 5000 people. For a project on how to develop acoustic models for speech recognition, Giampero Salvi at KTH trained HMMs on the Swedish SpeechDat database (Salvi, 1999), which he kindly permitted for use in this project. The HMM is a monophone model (50 monophones) with 8 mixture components. It uses 39 MFCCs (including delta and delta deltas). However, this HMM is based on telephone speech and will not work properly on recordings containing higher frequency components. Therefore, when this HMM is applied to non telephone recordings, resampling of the recording to 8 kHz is done before aligning is attempted. Ongoing work includes training on high quality speech from the Waxholm database (Bertenstam et al., 1995) to be able to apply a different HMM to recordings with higher frequency components and also to be able to make a Swedish aligner within EasyAlign freely available to download. However, to be able to train on this database a lot of conversions of labeling files and sound files has been done with several different scripts, but so far without very successful results.

Test Material

In forensic speaker identification case work it is common to receive a tapped telephone recording together with an orthographic transcription. To save time it would be very helpful to get an at least crude alignment between the text and the speech to be able to orientate and analyze different parts described in the orthographic transcription. As test material in this project 26 seconds from an authentic recording of a tapped telephone conversation was used together with its orthographic transcription. The transcription is preprocessed by dividing the text into line chunks depending on punctuation marks or phrasing. The number of lines then decides how many parts that are being aligned, i.e. pausing and lines should coincide to optimize the alignment. The Swedish orthographic transcription used was chosen due to its uncontroversial forensic content (lines as divided by preprocessing):

Hallå.

Ja go morron go morron.

Tjena.

Hur är läget?

Nä, bra.

Är det bra?

Gott å komma hem.

Ja fan du har ju vart i fjällen va, det hade jag helt glömt bort vet du.

Jaha.

Jag satt och ringde och ringde och var så törstig va.

Ja precis, nä ja kom hem igår kväll då någon gång.

Jaså.

Ja.

Å fan, se där ja, haft kul i alla fall?

Jaja, har vart roligt som fan vet du.

Åkt bräda eller?

Ja, både och faktiskt.

Både och?

Procedure

Utterance segmentation was done with line breaks according to transcription above and then manually corrected not to mess up the alignment of the phonetic segments.

The textgrid was then converted into a phonetic string using the implemented G2P and then manually corrected before attempting to align on phone level using HVite and the HMM for phone speech described above.

Results and Discussion

First a crude utterance segmentation was done with the line breaking according to above. The result shows that it is possible to achieve reasonably accurate results just by using a low amplitude level to interpret silence, measure it and apply a threshold for pausing (here 90 milliseconds). Figure 1 below illustrates the result of the crude utterance segmentation.

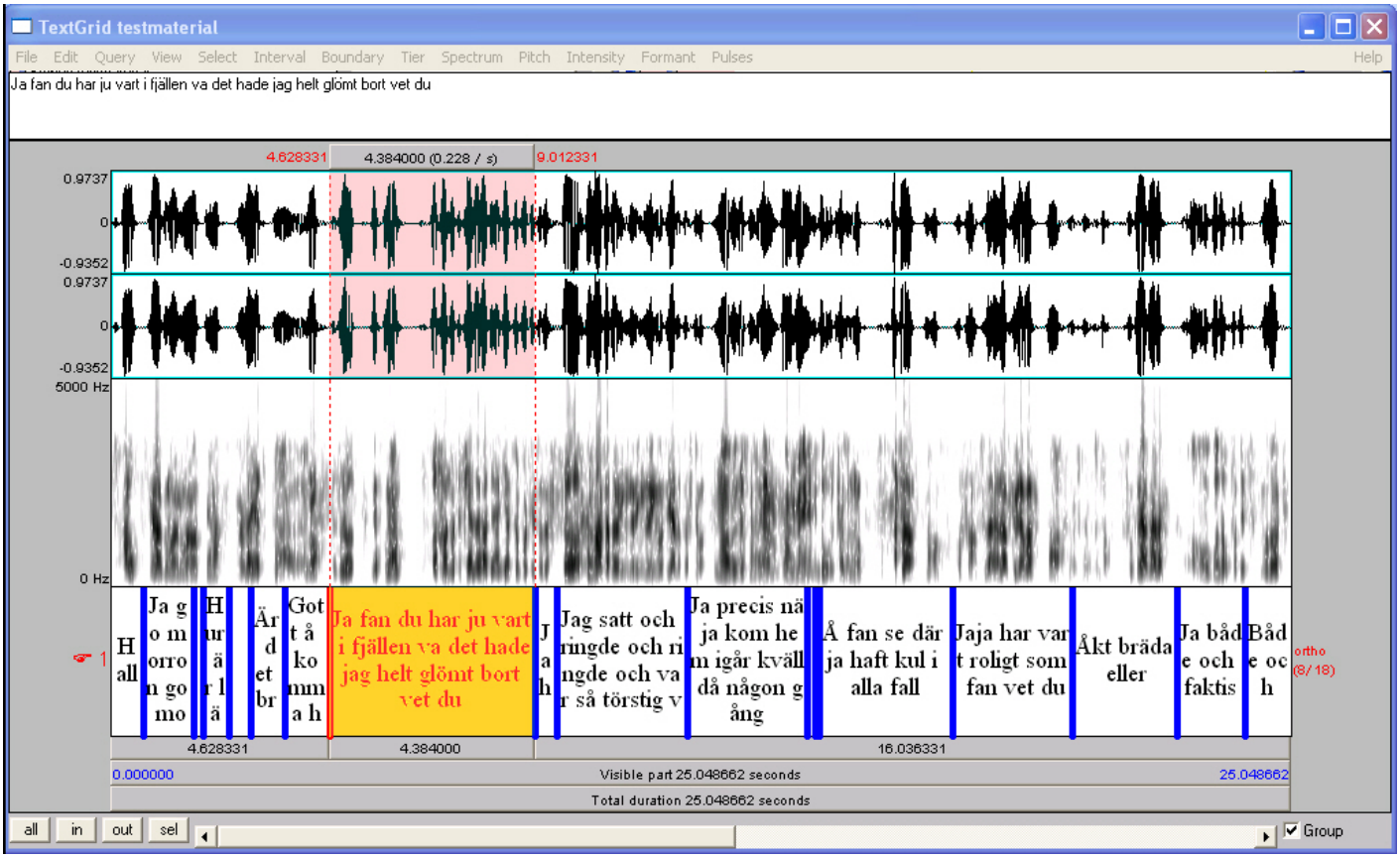

Figure 1. Result of the crude utterance segmentation.

A manual correction was then made of the utterance segmentation so that unnecessary mistakes would not be made by the segmentation at the phone level. However, first a crude conversion was made to a phonetic string. The result is

illustrated in figure 2 below.

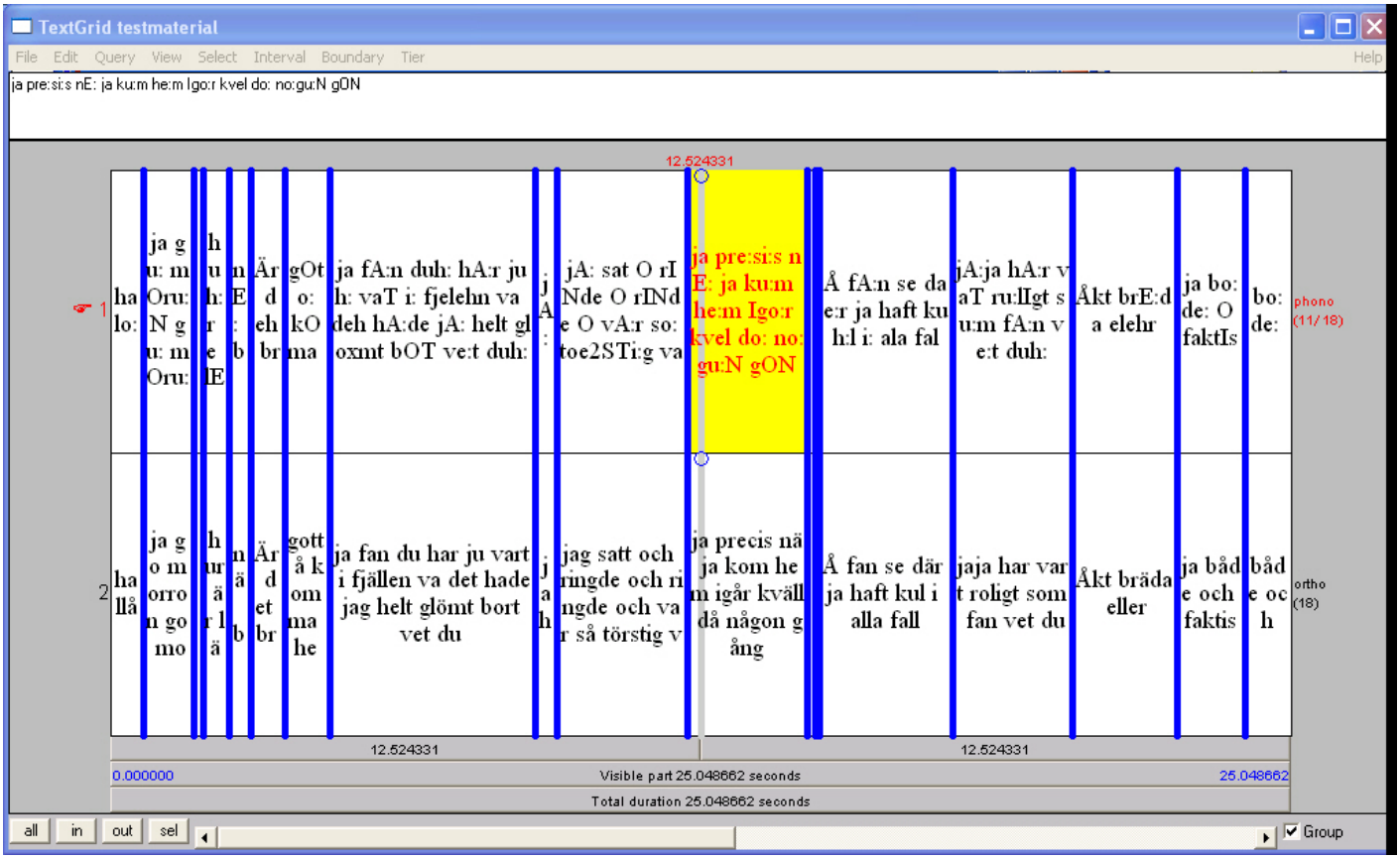

Figure 2. Result of the crude conversion into a phonetic string.

As the result is not completely satisfactory a quick manual correction was made of the phonetic transcription as well before an attempt to align the whole sequence into more detailed levels was pursued. In figure 3 below it is possible to see an overview of the result of the alignment on phone level.

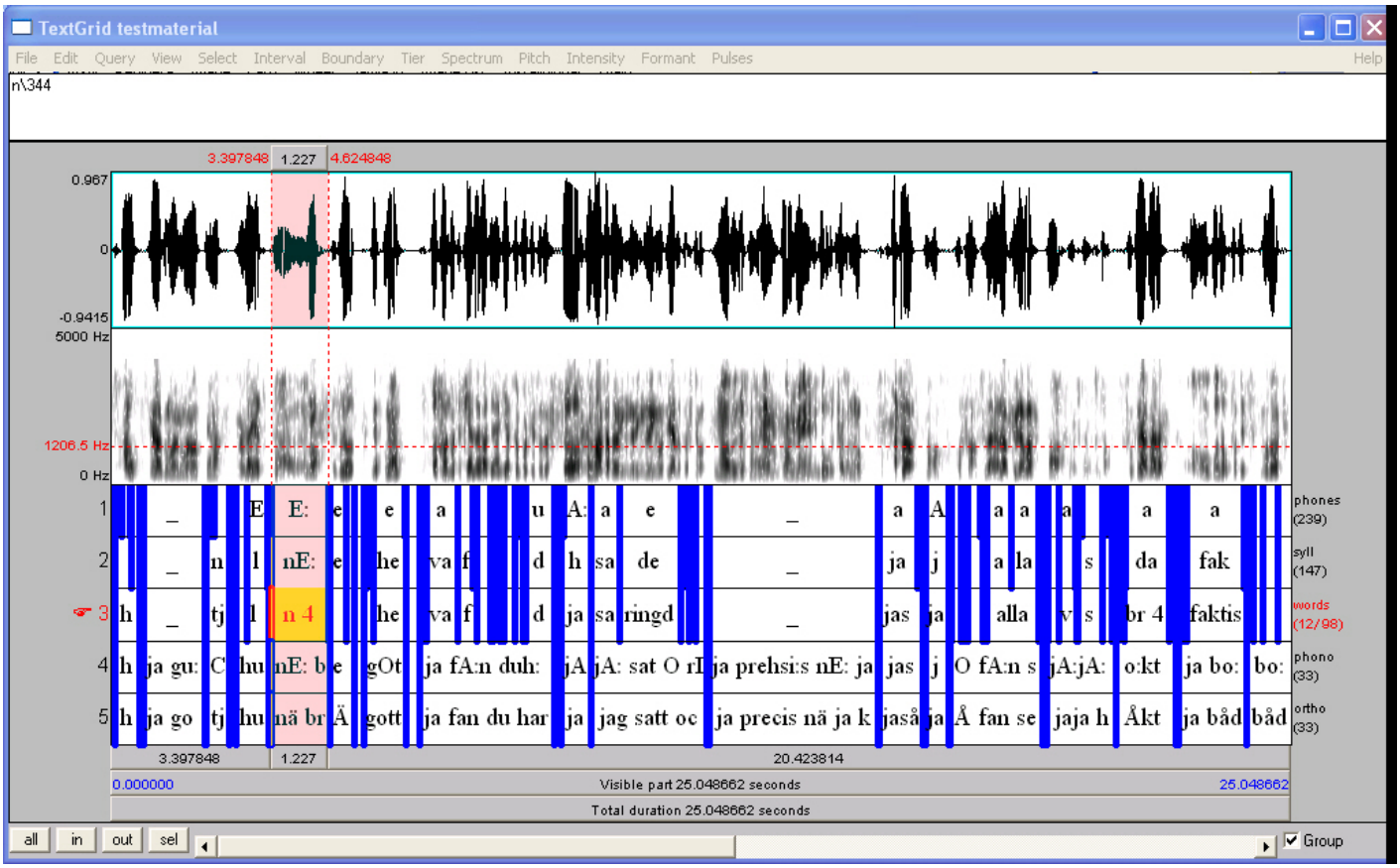

Figure 3. Overview of the result from the full aligning.

The precision at the phone level can easily be judged as very crude and quite unsuccessful, where vowels are aligned with the highest precision. At this stage in the process, some parts are not aligned at all due to some reason that is not fully investigated yet. At syllable and word level the aligning is much better and probably also much more useful.

Conclusions

This preliminary testing of implementing a Swedish semi-automatic aligning in Praat has been very successful as there now is an existing tool that can be used, even though the results at several instances are very crude and imprecise. Considering the quality of the tested material, the result is however very useful as a first step to facilitate the process of segmenting and analyzing recordings. When it comes to other kinds of databases with a lot of material, I think this is a enormous step forward from

not having anything time aligned with the signal at all.

Future Work

There are still some bugs existing in the transfer of some transcription segments and also interpreting some of the results from HVite. A log file is saved with warnings from HVite as well as errors from scripts are directly reported in an info window and bugs fixed with time. The future offers many developments. First of all the training of suitable material and bugs fixed must be done so a proper aligner can be evaluated comparing the output to a manually labeled material. Secondly, training will be possible directly from within Praat together with a Praat labeled database in the future and more languages added to EasyAlign. Adaptation to new speakers will then hopefully be implemented or added as a feature to make it possible to train on new material (for example forensic recordings). With EasyAlign as a tool, new frameworks for both speech and speaker recognition can be developed within Praat. These new frameworks are then suited for all kinds of analysis and displaying of data, results from investigation etc.

Acknowledgement

I would like to express my gratitude to Giampero Salvi and TMH (Speech, Music and Hearing) at KTH for letting me use an HMM trained on the SpeechDat corpus.

References

Bertenstam J., Blomberg M., Carlson R., Elenius K., Granström B., Gustafson J., Hunnicutt S., Högberg J., Lindell R., Neovius L., Nord L., Serpa-Leitao A., Ström N. (1995) Spoken dialogue data collected in the WAXHOLM project. *STL-QPSR 1/1995*, 49-74.

Boersma, P. & Weenink, D. (2007) Praat: doing phonetics by computer (Version 4.6) [Computer program]. Retrieved May 16, 2007, from http://www.praat.org/

Elenius, K., & Lindberg, J. (1997). SpeechDat - Speech databases for creation of voice driven teleservices. In Bannert, R., Heldner, M., Sullivan, K., & Wretling, P. (Eds.), Proceedings of Fonetik 1997, Dept of Phonetics, Phonum 4 (pp. 61-64). Lövånger/Umeå.

Goldman, J.-Ph. 2007. EasyAlign : a semi-automatic phonetic alignement tool under Praat. Submitted to *16th Int. Cong. Phon. Sc., Saarbrücken, Germany*, August 6-10, 2007.

Malfrère F., Deroo O., & Dutoit T., 1998 Phonetic alignment: Speech synthesis vs. hybrid HMM/ANN. *Proceedings of ICSLP '98*, 1571-1574.

Malfrère, F. et Dutoit, T. (2000) Alignement automatique du texte sur la parole et extraction de caractéristiques prosodiques, in *Ressources et évaluation en ingénierie des langues*, Chibout, Mariani, Masson, Néel ed., De Boeck et Larcier, Paris, pp. 541-552.

Salvi, G. (1999). Developing acoustic models for automatic speech recognition in Swedish. The European Student Journal of Language and Speech. http://www.essex.ac.uk/web-sls/index.html

Sjölander, K. (2001) Automatic alignment of phonetic segments. Working papers 49: Papers from Fonetik 2001, Lund, Lund University, Dept. of Linguistics, 140-143.

Sjölander, K. (2003) An HMM-based system for automatic segmentation and alignment of speech, *in* 'Proceedings of Fonetik 2003', pp. 93–96.

Young, S. et al. (2005) The HTK Book. 2006 (for version 3.4). Cambridge University Engineering Department: http://htk.eng.cam.ac.uk/docs/docs.shtml