Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models

1 AUTHOR:

Stefan Rapp
Conante

32 PUBLICATIONS 301 CITATIONS

SEE PROFILE

Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models

An Aligner for German

Stefan Rapp

Institut für Maschinelle Sprachverarbeitung (IMS),

Universität Stuttgart, Azenbergstr. 12, 70174 Stuttgart, Germany

e–mail: Stefan.Rapp@ims.uni-stuttgart.de

Abstract

This paper describes the architecture of a word and phoneme aligner based on Hidden Markov Models (HMMs). It was developed to allow for word, syllable and segment length extraction as part of a feature extraction stage for prosody recognition. From a given orthographic ASCII text and sampled speech data, a label file with phonemes, syllables or words is automatically generated. Linguistic categories coded in the lexicon can also be included in the generated label files.

1 Introduction

The task of phoneme alignment is closely related to the automatic speech recognition (ASR) task [Rab89, HTK94]. Roughly speaking, phoneme alignment is speech recognition without

a large portion of the search problem. In speech recognition, given recorded speech, we are interested in what words have been spoken and don't bother about the start and end points of the individual segments. For alignment, it's much the other way round. The string of spoken words is supposed to be known; the only thing of interest are the start and end points of each segment or other linguistic category to be aligned, e.g., syllable, stress foot, morpheme, phrase etc. By restricting the possible paths to just one string of categories, we greatly reduce the speech recognition search problem leaving time as the only degree of freedom. As the task of phoneme alignment can be considered as simplified speech recognition, it is natural to adopt a sucessful paradigm of ASR, namely HMMs [Rab89, HTK94], for alignment. In fact, in our system the training stage for the aligner does not differ from training for ASR. In the training stage, a stochastic model for every phoneme is calculated from recorded training data. For training as well as for the actual alignment, some preprocessing steps are necessary. The orthographic text is converted into a string of phonemes and the speech data has to be coded into feature vectors to maximize in-class similarity and between-class differences. In training, model parameters are estimated in an iterative process to fit best the training set of utterances. For alignment the trained phoneme models are concatenated and run against the feature vectors produced from the sampled speech data using Viterbi decoding [HTK94]. As a result, the most probable path through the models is found, giving the points in time for every transition from one phoneme model to the following.

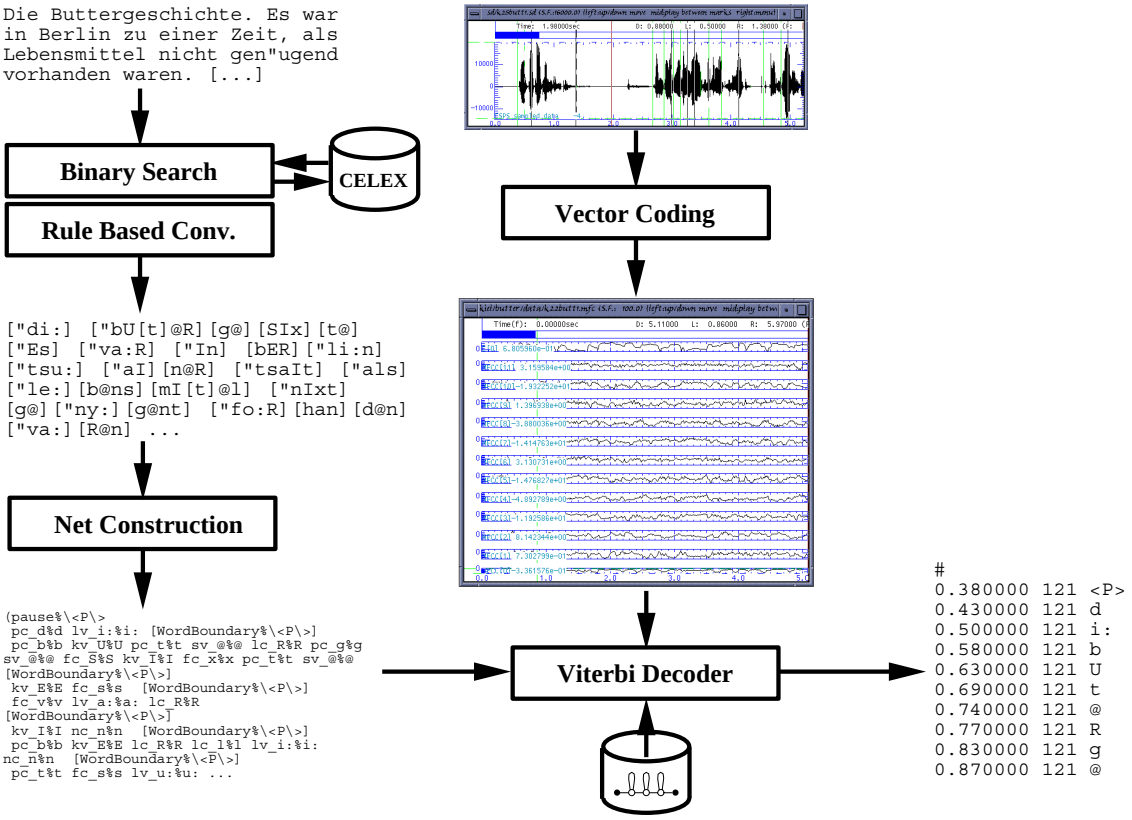

Figure 1: Architecture overview

2 Architecture

Our system was developed using version 1.5 of Entropic's Hidden Markov Toolkit [HTK94] and the 359 611 word forms of the German part of the CELEX lexicon [Baa&al93] for grapheme to phoneme conversion. Figure 1 shows the overall architecture of the aligner. The two inputs, orthographic text and sampled speech data (as shown on top of fig. 1), are preprocessed and then fed into the HTK Viterbi decoder which does the actual alignment. The next two subsections describe the two preprocessing stages for the orthographic text and speech data respectively.

2.1 Grapheme to Phoneme Conversion and Net Construction

Firstly, the text has to be converted into phonemes. In our system, this is done mainly by lexicon lookup. For the 359 611 word forms from CELEX, binary search has shown to be fast enough. Since for alignment the search problem is restricted to the time dimension, the rather large size of the lexicon does not decrease performance as it would be the case in ASR. Alignment should be able to handle unrestricted text, and so the system benefits from a large sized lexicon. A rule based text to phoneme conversion [Tab95] as a fall back strategy has also been incorporated in cases when lexicon lookup fails. Alternatively, it is possible to insert phonetic transcription directly into the orthographic text, for example in the case of foreign proper names which can neither be expected to be in the lexicon nor translated correctly by the german pronunciation rules implemented by the text to phoneme conversion. As all conversion stages use UNIX[1]-style IO leaving unprocessed text in the output, it is easy to switch or extend conversion procedures, e.g., for correct handling of numbers, dates, abbreviations etc. HTK depends on a regular grammar for the Viterbi decoder to prune the search space of possible word sequences for the ASR task. For alignment, the regular grammar is simply the linear concatenation of phonemes from the conversion stage and is generated from it by the use of standard UNIX tools. The only variability is given by inter word pauses. The grammar is generated in a way such that pauses may or may not be put between two words. As a consequence of these minor path selection possibilities, a language model does not have to be considered. Its only function would be to find the inter word pauses, a task which is done sufficiently well without a language model.

[1] UNIX is a trademark of AT&T.

2.2 Vector Coding

As a second input, Viterbi decoding needs vectors that describe the sampled speech data in terms of features. Our system uses spectral information by converting the speech data into twelve mel frequency cepstral coefficients plus overall energy. From these parameters delta and delta-delta values (corresponding to first and second derivatives) are calculated giving a total of 39 parameters per vector. Vectorframes have a distance of 10 ms, they are calculated using a Hamming weighted window of 25.6 ms length and a preemphasis of 0.97. The analysis order was chosen to be 24.

3 Training

Training was done using 200 phonetically balanced sentences from "The Kiel Corpus of Read Speech Vol. 1" [Koh94, Koh95], which were uttered by 12 male and female speakers (100 sentences per speaker). The training material sums up to 6144 spoken words. The HMMs were initialised using the elaborated phonetic transcription done in Kiel, but then the further training steps were done with the more phonologically oriented transcription found by lexicon lookup. Since for this transcription and the training data no correct alignment is available, the training was done using embedded reestimation [HTK94]. Note that this approach yields a rather implicit modeling, since the lexicon does not contain pronunciation variants which can be observed quite frequently in the training material. The explicit training of phonetic variants would also be possible, for our task description, however, it turned out to be unnecessary. It has been also claimed [Cam95], that phonetic variation can be derived from a phonologically aligned transcript in a systematic (i.e. rule-based) manner. Further experiments are necessary to clarify this issue.

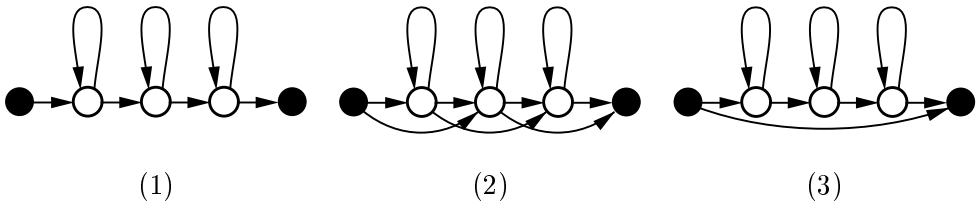

Figure 2: Model topology used for phonemes

3.1 Phoneme Model Topology

The phonemes are modeled by simple context independent left-to-right HMMs with 3 emitting states (4 emitting states for diphtongs), single mixture gaussian output probability density functions and a diagonal covariance matrix. We tried three different model topologies which are shown in fig. 2, (1)–(3). The models are (1) without skip transitions, (2) with three skip transitions, and (3) tee models, i.e., models that can be completely skipped. Surprisingly, the best results were achieved with the simplest of the models shown in fig. 2, i.e. model (1). When more freedom to the automatic procedure was given, more errors occured.

3.2 Modeling of Speech Pauses

Some problems arose with speech pauses, because breathing was not labeled in the speech data. Sequences of silence followed by breathing followed by silence were simply labeled with "pause". As a result, when modelling pauses with the simple model as shown in fig. 3 (1), only the silence before or after breathing was labeled as a speech pause with the automatic alignment procedure. Breathing and the other segment of silence collapsed with preceeding or following segments, especially plosives and fricatives. We tackeled this problem as follows. A more elaborate model as shown in fig. 3 (2) was choosen. The

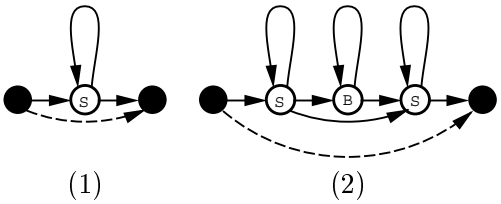

Figure 3: Model topology used for pauses

same output probability density functions (pdfs) as for the simple model from fig. 3 (1) were used as initialisation for the first and third state marked with "S" in fig. 3 (2). In addition, the output pdfs of these two states were tied, i.e., updated together in training. The output pdfs of the middle state (marked with a "B") were copied from the middle state of the fricative `[h]`. By the following embedded reestimations, the model adapted to model silence-breathing-silence sequences quite well, as evaluation on test material has shown.

3.3 Evaluation of Training

When reestimating the parameters of the HMMs, it is necessary to decide when to stop the iterative training process. Figure 4 shows the learning curves for different accuracy threshold values. The curve at the bottom refers to the percentage of segments that lie in between a 10ms region around the manual labelling. Above, the respective curves for accuracies of 15ms, 16ms, 20ms, 32ms, 48ms and 64ms are plotted. They were calculated using unseen speech data ("Die Buttergeschichte", [Koh95]) uttered by speakers that did not appear in the training set. As can be seen from the figure, the best point to stop training is right after the first iteration. For a 20ms accuracy (two frames), the procedure reaches 76.82%. This means that about one fourth of the automatically found segment boundaries mismatch the manual labelling by 20ms or more.

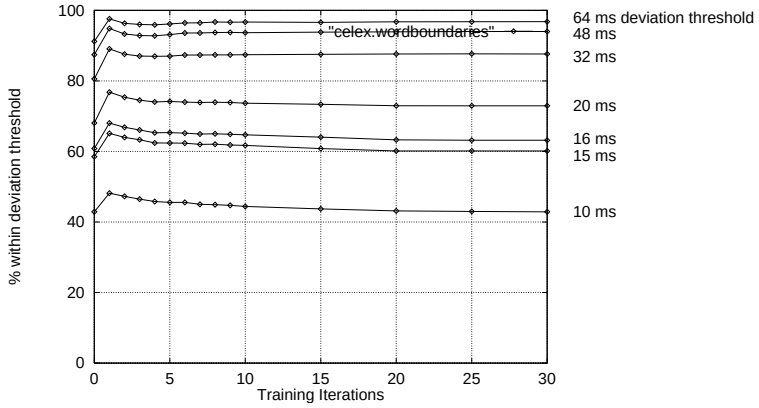

Figure 4: Percentage of wordboundaries within 10, 15, 16, 20, 32, 48, 64 ms from the manual labelling (curves from bottom to top resp.) over iteration cycles.

4 Alignment

At present, two versions of the aligner are used in our laboratory. One version (called `alignphones`) finds start and end points of phones if speech data and orthographic text are given. If the text is missing, the user is prompted for it while the speech data is played back. The other version, `alignwords`, finds start and end points of whole words if speech data and orthographic text are given. This is done by a different net grammar in which the Viterbi decoder produces a word label only for the last recognized phoneme of a word rather than a phoneme label for each recognized phoneme. As a result, a label file with words is produced compatible with the phoneme labels, that is, word boundaries coincide with the phone boundaries, and pause labels are identical in both files. Syllable labels are not produced from the aligner directly, because the Viterbi decoding with HTK can not handle overlapping segments. As many phonologists claim, e.g. [Venn90, Hall92], German has ambisyllabic consonants, i.e., consonants that count as a coda of a preceeding as well as an onset of a following syllable resulting in an overlap of the two syllables involved, the "Gelenk" in the terminology of Vennemann. However, we achieve explicit syllabic labelling

with our aligner because it is linked to the CELEX lexicon which gives syllabification (with ambisyllabic consonants) for each of its almost 360.000 phone strings. Similarly, our Aligner can provide morphological and morphosyntactic label files because these are included in the CELEX lexicon as well. The theoretical consistency and descriptive quality of these files, however, is totally dependent on lexicographic coding of these linguistic categories. In cases of polymorphy additional tools (morphological and syntactic taggers [Sch95]) can be used to disambiguate or at least improve category labels of the output string.

A remaining problem is the rather coarse time resolution given by the frame rate. Some experimentation with larger frame rates, 200 Hz resp. 1000 Hz instead of the current 100 Hz (that is 5 resp. 1 ms instead of 10 ms frame to frame distance) showed, that the robustness of the aligner decreases when attempting to increase accuracy, as fig. 5 shows. A frame rate of 10 ms performed best for all threshold values, but a frame rate of 5ms seems to be acceptable as well. Note that for 5ms and 1ms curves, embedded reestimation does not improve performance, as the values for 0 iterations are the highest ones. For 10 ms frame distance, best performance is achieved with one training iteration. The plotted data for 10 ms differs from the figure above, as for this figure transition probabilities had to be reestimated before the embedded reestimation cycles.

5 Conclusion

This work has shown that it is possible to build a robust phone and word aligner for German with easily available tools [HTK94] and resources [Koh95, Baa&al93]. The aligner can compete with comercial products and research tools from other departments developed for other languages, as fig. 6 shows. The use of a large lexicon allows good coverage of linguistic

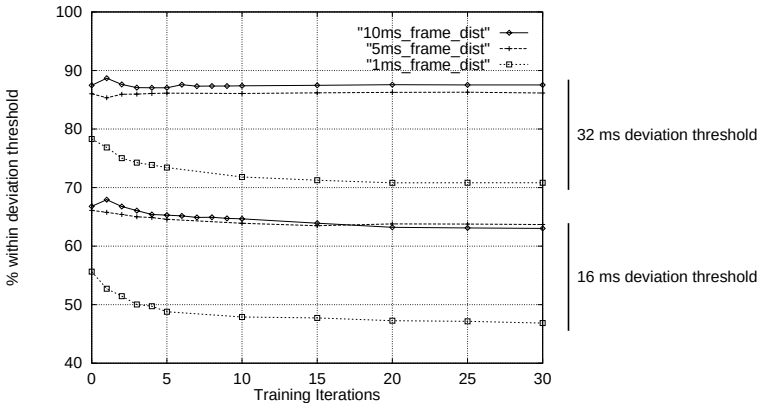

Figure 5: Percentage of boundaries within a given threshold for different frame rates. Upper lines correspond to a 32 ms threshold, lower lines to 16 ms.

forms. Moreover, the choice of a particular lexical implementation (CELEX) gives us the possibility to produce syllabified, morphologically and morphosyntactically labelled speech for known texts. Although the phonetic representation of CELEX is strongly phonology oriented, the procedure has shown to be quite robust even if the actual speech is erroneous or reduced. The tool is flexible and not totally dependent on the choosen lexicon, because the lexicon can be either replaced by another lexicon or even completely substituted by a rule based grapheme to phoneme conversion like [Tab95], if the quality of the phoneme transcription is less crucial. At present, a rather good estimation of syllable and word length is achieved, since for the larger units, the quantisation to 10ms frames is less crucial. Hence, the performance of our aligner is fully acceptable for the task it has been devised for, i.e. the recognition of syllable and word based prosodic phenomena. For fine phonetic labelling for which a time resolution under 10 to 20 ms is required, other alignment tools should be conceptualised. Beside possible uses for prosody recognition, the aligner is also being used for data-based concatenative speech synthesis of arbitrary voices in our lab [Kamm]. For prerecorded voices, the aligner does the required “cut”; and the “paste”

	Dalsgaard, Andersen, Barry (1991)			Kvale (1993)				Entropic's Aligner	This system a b (c)
	Danish/English/Italian			Danish/English/Italian/Norwegian				English	German
10 msec	48.7	62.6	34.2	70.3	66.2	64.2	68.1		53.9 59.1 (60.8)
15 msec	58.9	72.5	45.8	81.0	76.4	77.7	80.4		70.6 75.9 (77.9)
20 msec	65.5	77.5	52.0	86.1	82.3	84.5	86.4		80.9 84.4 (86.3)
16 msec								71	73.7 78.4 (80.4)
32 msec								90	91.3 91.1 (92.4)
48 msec									96.5 94.5 (95.4)
64 msec								97	98.4 96.0 (96.7)

Figure 6: Percentage of boundaries within a given threshold from the manual labelling for various systems. Percentages for column 1 and 2 are taken from [Kva93]. Column 3 is taken from [Ent94]. Results for this system were evaluated with all parameters fixed (10ms frame distance, 1 embedded reestimation cycle) on new material ("Nordwind und Sonne", [Koh95]). Column a: lexicon lookup as described in this paper, only word boundaries are considered. Column b: forced alignment with phonemes from manual labelling, all boundaries considered. Column c: same as b but pauses not considered.

is established by the particular synthesis application. Recent research on concatenative speech synthesis [BlCa95] shows that concatenation of aligner produced speech chunks leads to highly natural speech synthesis.

References

[BlCa95] Black, A. W. and Campbell, W. N. (1995)
Optimising selection of units from speech databases for concatenative synthesis, *Proceedings EUROSPEECH '95 Madrid* Vol. 1, pp. 581–584.

[Baa&al93] Baayen R. H., Piepenbrock R. and van Rijn H. (1993)
The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1993.

[Cam95] Campbell, W. N. (1995)
From read speech to real speech, *Proceedings ICPhS 95 Stockholm* Vol. 2, pp. 20–27.

[Ent94] Entropic Research Laboratory, Inc, (1994)
Aligner Automatic text-speech alignment software. (Promotion flyer) 600 Pennsylvania Avenue, Washington DC 20003.

[Hall92] Hall, T. A. (1992)
Syllable structure and syllable-related processes in German, Niemeyer, Tübingen.

[HTK94] Entropic Research Laboratory, Inc, (1993)
HTK – Hidden Markov Model Toolkit. 600 Pennsylvania Avenue, Washington DC 20003.

[Kamm] Kamm, C. (forthcoming)
Diplomarbeit Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

[Koh94] Kohler, K. J. (ed.) (1994)
Lexica of the Kiel PHONDAT Corpus, Read Speech, Volume I. *arbeitsberichte nr. 27/28 (AIPUK 27/28)*, Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel.

[Koh95] Kohler, K. J. (ed.) (1995)
The Kiel Corpus of Read Speech (CD-ROM). Institut für Phonetik und digitale Sprachverarbeitung, Christian–Albrechts–Universität zu Kiel, D–24098 Kiel, Germany, e–mail: `gph33@rz.uni-kiel.de`.

[Kva93] Kvale, K. (1993)
Segmentation and labelling of speech. Dissertation, The department of telecommunications of the Norwegian institute of technology.

[Rab89] Rabiner, L. R. (1989)
A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE,* Vol. 77, No. 2, pp. 257–285.

[Tab95] Tabar, E. (1995)
Graphem–Phonem Konvertierung und Silbentrennung für die deutsche Sprache. Studienarbeit Nr. 1474, Fakultät Informatik und Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

[Sch95] Schmid, H. (1995)
Improvements in Part–of–Speech Tagging with an Application to German. *Proceedings of EACL SIGDAT-Workshop*, Dublin, Ireland.

[Venn90] Vennemann, T. (1990)
Syllable structure and syllable cut prosodies in Modern Standard German. In: Bertinetto, P. M., Kentstowicz, M. & Loporcaro, M. (eds.) *Certamen Phonologicum II. Papers from the 1990 Cortona Phonology Meeting*, Rosenberg & Sellier, Turin 211–243.