Confronting HMM-based Phone Labelling with Human Evaluation of Speech Production

Jan Volín, Radek Skarnitzl*, Petr Pollák***

* Charles University in Prague, Faculty of Arts and Philosophy, Czech Republic
** Czech Technical University in Prague, Czech Republic

{jan.volin, radek.skarnitzl}@ff.cuni.cz, pollak@fel.cvut.cz

Abstract

The paper presents outcomes of an experimental study in which various modes of HMM labelling were tested on two groups of Czech speakers differing in the quality of their performance. Apart from the highest precision for the given speaking style - in this case read connected speech - we were also looking for indications that the HMM labeller might function differently for the group of good speakers and for the group of poor speakers. It turned out that impressionistically evaluated good and poor speakers were, at least in some of the modes, labelled with a different degree of precision.

1. Introduction

Phonetic research in recent years is to a great extent dependent on large corpora of speech. Speech databases have been successfully used to provide useful information on quite diverse types of problems. The common property of such databases is that the material in them is labelled. Manual labelling of speech items, however, can be a tedious job, and researchers have devoted a lot of effort to find at least semi-automatic ways of labelling corpora since it is generally accepted that the greater the pool of correctly labelled speech items, the more reliable findings it provides. In the field of segmental research, one usually requires correctly labelled boundaries of individual phones which represent phonemes of a given language. This task can be relatively successfully accomplished with various automatic speech recognition (ASR) methods, such as the standard HMM-based forced alignment algorithm (e.g., [1], [2], [3]).

It is clear, however, that characteristics of speech can change considerably as a function of different speaking styles (e.g., [4], [5]). Instead of looking for the best universal labelling tool then, it might seem quite practical to try to design the best labelling tool for a given speaking style. For that purpose, though, one has to know which speaking styles or modes are different enough to make use of a focused HMM procedure. There have been some useful and sometimes quite elaborate attempts to classify speaking styles [6], but it is still true that even within one style, human listeners can further classify the performance of a speaker with regard to proficiency and overall impression. The question is whether these differences are relevant to technologies used in ASR. In search for the answer, we decided to use read connected speech and investigate whether human assessment of the speaker's performance correlates with the success rate of several distinct HMM-based labelling modes.

The standard way of evaluating the performance of an automatic segmentation procedure is to compare the results with manually labelled data [7]. To this end, we used manually labelled paragraphs read by 25 good and 25 poor speakers, and a trained HMM-based recognizer was used to find boundaries of the known phones automatically. A detailed analysis of the automatic labelling precision, such as ours, is not available for the Czech language yet.

We hypothesised the following outcomes. First of all, as a null hypothesis, we could expect no differences between the good and poor speakers for any of the labelling algorithm settings used. At the same time, various algorithm settings would perform on the same level of precision. The opposite outcome would lead to higher precision of the labelling in case of one of the groups. Intuitively, we might expect good speakers to lead to greater success in correct placement of phone boundaries. Moreover, various algorithms used would differ from each other in their efficiency. All the other possible outcomes would be combinations of the previous two.

The ultimate question then would be whether it is sensible to consider an HMM labelling tool that could readjust its settings after receiving some sort of indication of a speech style in order to increase its success rate.

2. Method

2.1. Material

A group of 265 Czech university students were asked to read a short story of nine sentences as fluently and as naturally as possible; they were given adequate time for preparation. The recordings were made under identical conditions in a sound-proof booth with an electret microphone IMG ECM 2000 and a soundcard SB Audigy 2 ZS. The recordings were later presented via headphones to 6 evaluators who were asked to assess the performance of the speakers by awarding a mark from 1 to 4 based on the overall impression of speaker efficiency. Mark 1 meant an excellent speaker, mark 2 a very good speaker, mark 3 a just tolerable speaker and mark 4 a clearly inferior speaker.

Naturally, such marks are not parametric measures, which bears on statistical methods that can be used to process them. As a crude measure of a person's performance, however, the average score across the 6 marks awarded by the assessors was accepted. Cases with low homogeneity of opinion (differing by 2 grades) were excluded from further processing.

The arithmetic mean across all the scores awarded was 2.39. Thus, the scores of 'average speakers' oscillated around this value. We chose arbitrarily a band of 0.6 points on both sides of the average score to find good and poor speakers. Those with scores equal to or worse than 3.00 were classified as 'poor speakers', while those with scores equal to or better

than 1.78 were classified as 'good speakers'. We ended up with 25 good and 25 poor speakers whose recordings were manually labelled by two experienced labellers as a reference for success rate analysis. The whole material consisted of about 2,500 words and 11,000 phones.

2.2. HMM labelling setup

We used HTK toolkit [8] to build an HMM forced alignment algorithm. In this section we give a more detailed description of the settings of the labelling procedure, especially the parameters whose influence was analyzed in the present study.

2.2.1. HMM model structure

A standard five-state model for Czech monophones with three emitting states joined by two additional non-speech models (for silence and short pause) were used. Obviously, context independent monophones must be considered for the purpose of finding phone boundaries. All speech models were left-right models without any possible skips over a state.

It is well known that this three-emitting-state structure may cause some problems in fast fluent speech, i.e., when the realization of the phone is shorter than the minimum duration corresponding to the given three states. To analyze the effect of this phenomenon was also one of the tasks of our study.

Our initial experiments used HMMs with just one stream and no mixtures, but it is known that modelling can be considerably improved by using more mixtures of Gaussian functions describing the respective states, as well as more streams [9]. That is why we also used models with three streams and 32 mixtures, so as to analyze the effect of this more sophisticated modelling.

2.2.2. HMM models training

The models were trained by the standard Baum-Welch re-estimation algorithm. As the target data were not sufficient for training, the training procedure was conducted on two other large databases: SpeechDat, a telephony database with signals sampled at 8 kHz, and SPEECON, with data sampled at 16 kHz. The target signals to be labelled had a sampling frequency of 22.05 kHz, and they were adequately down-sampled. The mismatch between the training set and the target data was not critical, since the target data did not contain any specific background noise.

2.2.3. Phone inventory

The inventory of phones was determined by requirements of ASR algorithms and originates from the Czech SAMPA. It does not contain certain specific allophones, e.g, it does not distinguish between the voiced [r̬] and voiceless [r̥], or between [x] and its voiced allophone. [ə], which is not a phoneme in Czech but resulted from phonetic reduction, was modelled as [e], and the glottal stop had to be modelled as a short pause. These distinctions are not crucial for ASR.

2.2.4. Speech parameterization

The five settings used in our experiments worked with two different parameterizations. First, it was standard mel-frequency cepstral coefficients with energy, delta, and delta-delta coefficients, known from the HTK toolbox as MFCC_E_D_A . Second, we also tested a parameterization based on PLP cepstral coefficients, which entails a pre-processing stage eliminating additive noise. One of the aims of our experiment was to check the contribution of such a robust parameterization with respect to the mismatch between the training and target data. PLP cepstral coefficients with pre-processing eliminating additive noise were evaluated by the tool *CtuCopy* [10].

The precision of automated labelling is influenced by the analysis frame length and its time step. We used two frame lengths of 32 and 16 ms, in both cases with a 50% overlap. Standard weighting by Hamming window and pre-emphasis of 0.97 were used in our parameterization. Table 5 summarizes the five algorithm settings used in the present study.

Abbrev.	Parameterization	f_s [kHz]	Segments	Mixtures
mfcc8o	MFCC_E_D_A	8	32 / 16	No
mfcc8x	MFCC_E_D_A	8	32 / 16	Yes
mfcc16o	MFCC_E_D_A	16	16 / 8	No
mfcc16x	MFCC_E_D_A	16	16 / 8	Yes
explp8o	EXPLP_E_D_A	8	32 / 16	No

Table 1: Overview of the five algorithm settings

3. Results

The results of automatic labelling were analyzed in terms of the five algorithm settings, as well as in terms of the success rate of labelling individual classes of phones. In the following paragraphs and tables we focus on vowels (V), consonants (C), sonorants (sonor.), voiced obstruents (+voice), voiceless obstruents (-voice), and breaks (br.), which include pauses, hesitation sounds, and glottal stops (see Section 2.2.3.).

3.1. Overall precision

Table 2 shows the deviations of automatic segmentation from human labelling across all five algorithm settings. The mean error for all segments was 26.4 milliseconds. The results suggest that consonant onsets tended to be labelled with greater accuracy than vowel onsets, but this difference is not significant. Within the consonant group, the boundary of sonorants was placed with significantly lower accuracy than the boundary of obstruents ($p < 0.005$). The difference between voiced and voiceless obstruents was negligible ($p = 0.38$). The last column in Table 2 corresponds to breaks. The breaks largely contribute to the overall error; if we calculate the deviations only for the segment classes, the total mean error drops to 23.1 ms.

	total	V	C	sonor.	+voice	-voice	br.
mean [ms]	26.4	24.7	23.7	27.6	21.7	21.1	53.8

Table 2: Mean errors in labelling across all five labelling modes used

3.2. Comparison of individual labelling modes

Table 3 shows the mean differences between the human labelling and the *explp8o* mode. We can see that the performance of this algorithm is rather low when compared with the average values in Table 2. Also the relationships

between the individual phone groups are slightly different - the precision is lower with consonants than with vowels, and the difference between voiced and voiceless obstruents is more pronounced.

	total	V	C	sonor.	+voice	-voice	br.
mean [ms]	36.6	34.0	36.3	43.3	35.5	29.4	52.4
SD [ms]	3.1	3.0	2.6	3.5	4.8	3.1	18.8

Table 3: Labelling errors of the *explp8o* mode

Table 4 below gives the results for the *mfcc8o* setting. There is a marked deterioration in the placement of the break onsets (br.), which accounts for the high total mean value. In general, the algorithm based on mel-frequency cepstral coefficients is more accurate than the one based on perceptual linear prediction cepstral coefficients.

	total	V	C	sonor.	+voice	-voice	br.
mean [ms]	25.8	23.4	20.9	23.1	18.8	21.0	72.9
SD [ms]	9.6	9.4	10.3	18.5	7.2	9.0	18.0

Table 4: Labelling errors of the *mfcc8o* mode

The figures in Table 5 suggest that the incorporation of mixtures leads to higher accuracy ($p < 0.005$). The relationships between the individual phone groups remain the same as with the no-mixture mode.

	total	V	C	sonor.	+voice	-voice	br.
mean [ms]	23.2	21.3	17.9	19.0	15.7	18.8	71.1
SD [ms]	5.1	5.3	4.4	9.0	5.4	3.9	23.5

Table 5: Labelling errors of the *mfcc8x* mode

The following two tables show results based on models trained on 16-kHz data (SPEECON database). Moreover, the analysis window is 16 ms, with a time step of 8 ms.

	total	V	C	sonor.	+voice	-voice	br.
mean [ms]	26.9	26.7	25.2	30.3	21.9	21.6	39.3
SD [ms]	3.7	3.8	3.7	6.4	5.1	2.8	13.1

Table 6: Labelling errors of the *mfcc16o* mode

Table 6 above gives the results of segmentation by the *mfcc16o* mode. First of all, there is a notable increase in accuracy for the detection of breaks. As for the accuracy of the individual segment classes, however, this algorithm setting has the highest mean error of all those based on mel-frequency cepstral coefficients.

	total	V	C	sonor.	+voice	-voice	br.
mean [ms]	19.3	18.1	18.2	22.3	16.8	14.7	33.5
SD [ms]	2.6	2.8	2.5	3.6	3.5	2.8	12.9

Table 7: Labelling errors of the *mfcc16x* mode

Table 7 indicates that the *mfcc16x* algorithm is the most successful one. The mean error for the breaks remains low, but there is also a significant improvement in the accuracy of finding the onsets of speech segments.

3.3. Good *vs.* poor speakers

As another point of interest, we wanted to see whether the algorithms would be more successful in segmenting speech of good speakers than that of poor speakers. The results are depicted in Figure 1.

A series of t-tests revealed that the segment onsets obtained by the *explp8o* algorithm did not show different mean errors for the two groups ($p = 0.23$). However, the algorithms based on mel-frequency cepstral coefficients produced differences which were significant or marginally significant ($0.01 \leq p \geq 0.08$).

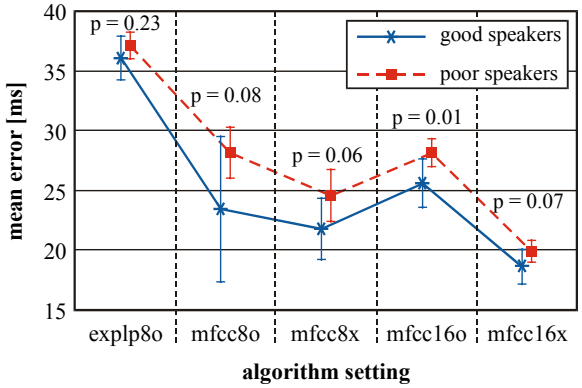

Figure 1: Diagram illustrating the difference between good and poor speakers across the five algorithms.

3.4. Phone duration effect

Working with 3-state HMM models with a prohibited skip over any of the states means that a phone can only be localized within the span given by the analysis frame and its overlap. The comparison of phone durations of manually and automatically labelled data was therefore also an important part of this study.

First of all, we investigated actual durations of the real phones, i.e., durations computed from the set of manually labelled data. Even though this analysis was based on a limited set of data (11,000 phones), the results confirmed durational relations among Czech phones in fluent speech. Table 8 gives duration means and standard deviations in ms of the longest and shortest monophones in our data set.

shortest phones			longest phones		
	duration	SD		duration	SD
[r]	46	18	[x]	114	68
[j]	50	16	[e:]	122	29
[ə]	50	26	[a:]	131	37
[d]	54	19	[au]	148	24
[l]	54	20	[eu]	164	24

Table 8: The longest and shortest phones (in ms) from the target data set.

The most varying phones with respect to duration were the fricatives [x] ($\bar{x}$ = 114 ms; *SD* = 68 ms) and [f] ($\bar{x}$ = 102 ms; *SD* = 55 ms), as well as some vowel sounds, short and long.

The duration means in Table 8 suggest that we might expect better results for the setting 16/8, which provides for the minimum localized phone duration of 24 ms, while the 32/16 mode requires the minimum phone duration of 48 ms.

Figure 2 illustrates the typical errors of the *mfcc8o* algorithm setting (the 32/16 mode) in determining the duration of phones. [l] and [a:] have been chosen as representatives of short and long phones, respectively (*cf.* Table 8). It is obvious from the histograms that the duration of the short [l] is determined with considerably lower accuracy with respect to the reference manual labelling.

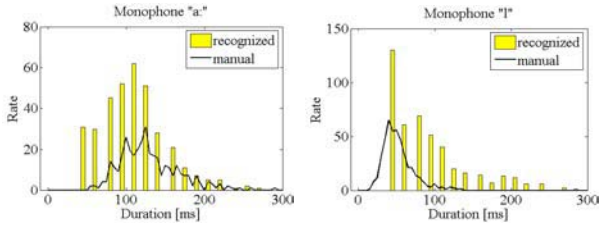

Figure 2: Phone duration with the *mfcc8o* HMM model.

In Fig. 3, we can observe a better fit for both histograms. This can be accounted for by the 16-kHz sampling frequency, as well as the shorter frame window and time step. The improvement is greater for the long [a:]. Slight improvement may also be observed for short phones representative [l] but some bias is still present due to the minimal duration requirement.

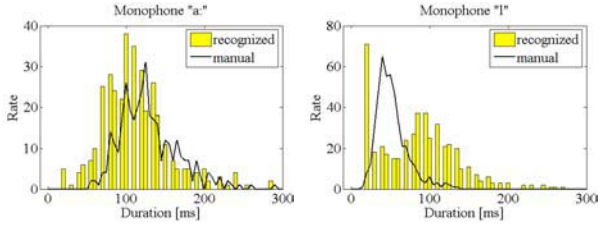

Figure 3: Phone duration with the *mfcc16x* HMM model.

4. Discussion

As far as speaker quality is concerned, impressionistic evaluation of speakers as good or poor is not based solely on segmental characteristics of their speech. It is influenced by prosodic features, as well as their voice timbre. Nevertheless, we found clearly significant effect of speaker quality when we used the *mfcc16o* paradigm for phone boundary detection. Even the settings *mfcc8x*, *mfcc16x*, and *mfcc8o* produced marginally significant results. This indicates that the problem deserves further research. One of the possible reasons why the effect of speaker quality was not stronger might be the fact that the two large training databases comprised supposedly both better and worse speakers.

As to the precision of labelling, the *mfcc16x* setting proved to be the best. One of the ways to achieve even greater precision might be to manipulate the overlap of analysis frames. We would be reluctant, however, to shorten the frame itself since this would lead to distortion of signal descriptors.

Other improvements can be sought in examining the mean errors of individual classes of segments. We found out that the main culprit here was the class *br*. (breaks), namely the behaviour of the glottal stop, for which the original models were not trained. We tried to model it as a short pause, but this approach did not bring satisfactory result. It is clear that a dedicated trained model of the glottal stop is necessary.

5. Conclusions

Our detailed comparison of manual and automated labelling performed on two distinct groups of speakers revealed that the technique based on forced alignment of trained HMMs is sensitive to a factor of speaker quality as perceived by human listeners. It also showed that for the Czech language it is necessary to model the glottal stop as an independent speech segment. Its occurrence is quite high and ignoring it leads to higher error rate even for labels of surrounding phones.

The analysis of phone durations suggests that for the speech style analyzed in our study it is necessary to look for even finer temporal resolution, especially for the sake of approximants and the central mid reduced vowel.

The presented technique seems to be a good tool with precision sufficient for at least basic pre-labelling of larger phonetic corpora, as well as for application in some other areas of speech processing.

In future, we would like to follow two lines of research: a) further improvement in the precision of our labelling procedure; b) investigation of various speaker-bound factors on the labelling process. We would also like to test the possibility of feeding the algorithms not with the list of exact phones present, but with the ideal text transcription.

6. Acknowledgements

This research was supported by VZ MSM 0021620825 to J. Volin and R. Skarnitzl, and by GACR 102/05/0278 and MSM 6840770014 to P. Pollak.

7. References

[1] Malfrere F., Deroo O., Dutoit T., Ris C., "Phonetic alignment: speech synthesis-based vs. Viterbi-based", *Speech Communication 40*, *2003*, *503-515*.

[2] Mella O., Fohr D., "Semi-automatic Phonetic Labelling of large corpora", In *Proc. of Eurospeech'99*, 1999.

[3] Nouza J., Myslivec, M., "Methods and application of phonetic label alignment in speech", *Radioengineering 9, 2000, 1-7*.

[4] Kohler, K. J., "Articulatory reduction in different speaking styles", *ICPhS 95 Proc*., Vol. 2: 12-19, Stockholm, 1995.

[5] Barry, W. J., "Phonetics and phonology of speaking styles", *ICPhS 95 Proc*., Vol. 2: 4-10, Stockholm, 1995.

[6] Eskanazi, M., "Trends in speaking styles research", *Proceedings of Eurospeech 93*, pp. 510-509, Berlin, 1993

[7] Pauws, S., Kamp, I., Willems, L. "A hierarchical method of automatic speech segmentation for synthesis application", *Speech Communication 19*, *1996*, *207-220*.

[8] Young S., et al, "The HTK Book (for HTK version 3.1)", Cambridge University, 2001.

[9] Boril H., "Recognition of Speech Under Lombard Effect", In *Proceed. of Speech Processing*, Prague, 2004.

[10] Fousek, P., Pollak P., "Additive Noise and Channel Distortion Robust Parameterization Tool – Performance Evaluation on Aurora 2 & 3", In *Proceedings of Eurospeech'03*, Geneve, 2003.