ESTIMATING THE QUALITY OF PHONETIC TRANSCRIPTIONS AND SEGMENTATIONS OF SPEECH SIGNALS

Maria-Barbara Wesenick, Andreas Kipp

Institut für Phonetik und Sprachliche Kommunikation (IPSK),
Ludwig-Maximilians-Universität München, Schellingstraße 3/II,
80799 München, Germany
wesenick|kip@phonetik.uni-muenchen.de

ABSTRACT

Our approach to the problem of evaluating segmentations and transcriptions of speech data is presented. We developed an automatic pattern-matching procedure that relates different manual or automatic segmentations to each other. The comparison of segmentations refers to the degree of identity concerning the chosen labels and of identity of segment boundaries. As we exemplify our evaluation method on the basis of automatic transcriptions of the Munich AUtomatic Segmentation System (MAUS) that is currently being developed at the IPSK (Kipp et al. [4]) our data also give information on the quality of the system's segmentation and transcription performance.

1. INTRODUCTION

For phonetic and phonologic investigations as well as for many applications in speech technology a large amount of segmented and labelled speech data are required e.g. for training and testing purposes in ASR. Already a few years ago with the creation of speech databases that contained manually labelled speech data the problem arose of how to estimate the quality of the available segmentations (Eisen et al. [3]). This question becomes even more important now that automatic segmentation systems are used. Aiming at an estimation of the transcription and segmentation performance of our MAUS system we came across the general problem of evaluating the quality of transcriptions. There is no straightforward method to make differentiated statements about the quality of segmentations apart from the assessment of spot checks, which does not allow a quantitative overall evaluation.

Our approach to obtain a telling quantitative description of the quality of speech segmentation and labelling is presented in this paper. It involves the definition of a reference to which segmentation data can be related. To avoid the problem of choosing arbitrarily one of the manual transcriptions as reference a one-leave-out-comparison is made. In this way we get as reference the overall judgement of the transcribers taken together. For the evaluation of the automatic segmentations the weighted manual segmentations serve as reference. The performance of MAUS i.e. the automatic segmentations are related to data that have been obtained manually and judged on the basis of observed deviation among manual transcriptions and segmentations.

The evaluation refers to two levels of congruity:

- identity of segment labels (of consonants)
- identity of boundaries of identically labeled segments

As a result of our study we obtain statistic data that enable us to make valid and precise statements about the quality of segmentations. The quality of an automatic segmentation system can be determined relating its performance to results achieved by human experts. Determining its weaknesses exactly makes specific improvements of the automatic system possible.

2. EVALUATION METHOD

2.1. Definition of a reference

A main problem evaluating segmentation data is to find an appropriate reference segmentation that can serve as a basis for the comparison of manual and automatic transcriptions.

We accept that conscientiously and sophisticatedly obtained manual segmentations are a reliable reference. But manual segmentations of speech signals are subject to individual judgements of human transcribers and therefore always differ to a certain degree without however in terms of one being more or less "correct" than the others (refer also to Eisen [2] and Eisen et al. [3]). This kind of differences are due to the fact that abstract and discrete phonetic symbols, that stand for distinctive units of speech, have to be related to the continuous and concrete acoustic speech signal. The consequence for an evaluation of an automatic transcription is that deviations from a single reference manual transcription can not necessarily be regarded as mistakes in the evaluated transcription.

Our evaluation therefore does not ground on a single selected reference segmentation, but on data derived by a group of human segmenters which serve as a measure for automatic segmentations. In a first investigation we determined the differences among the transcribers a) regarding the transcription of consonants and b) regarding the deviation of boundaries of identically labelled segments using a one-leave-out procedure.

The amount and kind of confusion of chosen labels and the degree of deviation of segment boundaries of different sounds and sound classes that is observed in manual analyses is chosen as allowed deviation of the automatic transcription from the reference.

Our aim in improving the performance of the Munich AUtomatic Segmentation System is to obtain results that show similar patterns of deviation to a similar extent compared to the weighted results obtained by human experts.

2.2. Speech Data

We used speech data of 10 speakers from the Phondat-II database of German (Pompino-Marschall (ed.) [5]) for which we have an automatic transcription generated by the MAUS system and manual transcriptions of 10 transcribers altogether. The data of one speaker, which comprise 64 read sentences, have been segmented in average by 3 human transcribers and in addition automatically by the MAUS system.

The level of labelling is broad-phonetic[1]. The labelling symbols are taken from the SAM-phonetic alphabet, providing a symbol for each phoneme of German. The manual transcriptions were done by relating an observed pronunciation to the citation form of the concerning word. The citation forms were given guiding the transcription.

2.3. Procedure

For the evaluation of manually obtained data two manual segmentations A and B of the speech data of one speaker have been related at a time. This has been done by a dynamic-programming pattern-matching algorithm, which computes the optimal match between the transcription symbols contained in A and B. Differences are classified as elisions (symbol x occurs in A but not in B), insertions (symbol x occurs in B but not in A), and replacements (for a symbol x of A a symbol y is found in B). In the one-leave-out comparison each segmentation serves as a reference for all of the other segmentations. So if e.g. four segmentations A - D are compared, the following combinations are looked at:.

AB	AC	AD
BA	BC	BD
CA	CB	CD
DA	DB	DC

The result of this comparison gives values for the overall agreement among human transcribers.

For the evaluation of the automatic transcription X it is compared to the manual transcriptions A - D. The data are obtained by comparing the following relations:

XA XB XC XD

The results from this second comparison are related to the overall agreement among manual transcribers and give information on how close the results are to our aim of obtaining reliable automatic segmentation and transcription data.

1. In our terminology we follow Barry/Fourcin [1].

3. CORRESPONDENCE OF LABELS

Table 1 shows in detail the data for the correspondence of chosen consonant labels (in%) among human segmenters (a) and the correspondence of labels of automatic transcriptions and the defined reference (b). Data for each label are shown as well as for the sound classes of stops, fricatives and nasals in addition to the values of an overall correspondence of chosen consonant labels.

	labels	a) manual transcriptions	b) automatic transcriptions
Stops	p	93.8	76.4
	b	97.8	82.5
	t	92.5	80.2
	d	79.6	75.1
	k	92.1	89.2
	g	85.9	72.1
	Q	86.6	78.3
	all stops	**89.9**	**80.2**
Fricatives	f	99.2	99.6
	v	96.5	88.2
	s	98.5	95.1
	z	92.9	98.6
	S	99.2	94.0
	C	98.3	94.3
	j	96.4	97.5
	x	99.4	92.9
	h	92.3	71.5
	all fric.	**98.0**	**93.6**
Nasals	m	98.2	97
	n	97.9	94.9
	N	93.4	83.5
	all nas.	**97.5**	**94.4**
	l	98.0	64.1
	r	96.0	99.0
all consonants		**94.8**	**88.4**

Table 1: Percentage of identical consonant labels among manual transcriptions (a) and of automatic transcriptions in relation to the defined reference (b).

It can be seen that in both types of transcription the identity of labels for stops (man: 89.9%, autom.: 80.2%) is lower than for fricatives (man.: 98.0%, autom.: 93.6%) and nasals (man.: 97.5%, autom.: 94.4%). (Similar results have been obtained by Eisen [2]). The difference between data for manual and automatic transcriptions furthermore is greatest regarding stops (9.7%), whereas it is smaller for fricatives (4.4%) and nasals (3.1%). The overall identity of consonant labels of manual transcriptions is 94.8%, the percentage of identical labels in automatic transcriptions and the reference is 88.4%, i.e. a difference of 6.4%.

A confusion matrix, that can be set up with the results from the dp-match, shows in which way consonant labels differ that have not been chosen identically. The data of table 2 show frequent consonant confusions. (They make up more than 0.8% of all labels for a concerning segment.). It is interesting to see that the confusion that occurs most often is in manual transcription the substitution of a voiceless plosive for a voiced one, respectively a voiced for a voiceless one (e.g. `p - b, t - d, k - g`). The same, only to a greater extent, has been observed in automatic transcriptions. It is worth noting that the kind of confusion in manual and automatic transcriptions is very similar. It can be seen that apart from the voiced-voiceless confusion mainly confusion in manner of articulation (b - v) or place of articulation (d - b, N - n) occurs. Also a combination of the confusion of manner and place of articulation and the voice/voiceless feature is observed (p - v, d - Q). It never happened, not even in the automatic transcriptions, that parameters of manner and place of articulation (e.g. p - s, N - t) have been mixed.

labels	a) manual transcriptions	b) automatic transcriptions
p	b - 5.2% v - 2.6%	b - 22%
b	p - 0.8%	p - 9% v - 4%
t	d - 4.4%	d - 10%
d	t - 11.4%	t - 10% p - 2% b - 2% Q - 1%
k	g - 5.6%	g - 5% Q - 1%
g	k - 8.7%	k - 11%
v	f - 2.1% b - 10%	f - 9.5%
s	z - 0.8%	z - 3.3%
N	n - 4.3%	n - 4%

Table 2: Confusions of labels (occurrence ≥ 0.8% of labels for the concerning segment) in manual transcriptions (a) and in automatic transcriptions (b) in relation to the defined reference.

4. DETERMINATION OF SEGMENT BOUNDARIES

In addition to the information on the degree of identity of chosen labels important information about the reliability of segmentation and transcription data can be obtained by looking at segment boundaries. Our investigation refers only to those segments that have been labelled identically.

.

symbol	sound class
Pvl	voiceless plosive
Pvd	voiced plosive
Fvl	voiceless fricative
Fvd	voiced fricative
L	lateral
N	nasal
V	vowel

Table 3: Phonetic Categories

segment boundary	a) manual segmentations	b) automatic segmentations
N-N	16	43
N-Fvd	11	34
V-L	14	36
V-Fvd	9	31
Fvl-Fvd	12	28
Fvl-Pvl	5	21
Fvl-Pvd	7	19
V-N	9	19
N-V	8	18
L-V	8	17
V-Pvd	12	19
Pvd-V	6	12
V-V	15	20
N-Pvd	11	15
Fvl-Fvl	11	13
Fvl-N	13	14
V-Fvl	7	8
N-Fvl	6	7
Fvd-V	12	12
N-Pvl	10	10
Pvd-N	9	9
V-Pvl	12	11
Pvl-Fvl	11	10
Fvl-V	7	6
Pvl-N	12	7

Table 4: Average deviation of segment boundaries of identically labelled segments in ms for manual segmentations (a) and automatic segmentations (b)

We referred to segment boundaries in terms of the classes of adjacent segments (see table [3]). The average differences of chosen segment boundaries in manual (a) and automatic (b) segmentations are given in Table 4. (The beginning and end of utterances are not regarded, neither are boundaries that occurred less than 0.3% of all transitions.) They are ordered according to the difference between manual and automatic segmentations in relation to the respective reference.

A general tendency can be observed that a boundary that is difficult to find for humans is also less correctly determined by the automatic system, e.g. the nasal-nasal and vowel-lateral transitions. Certain transitions are very well detected by the automatic system, e.g. voiceless plosive - nasal, voiceless plosive - vowel transitions.

Another way of describing the segmentation performance is to determine the amount of boundaries that lie within a certain time range. Table 5 shows the data for manual (a) and automatic (b) segmentations referring to different time ranges. It can be seen that as much as 61% of automatically determined segment boundaries are accurate within a range of 10 ms, which makes a difference of 26% compared to manual transcriptions.

time range	a) manual segmentations	b) automatic segmentations
= 0 ms	63%	1% (<0.5ms: 15%)
< 5 ms	73%	36%
< 10 ms	87%	61%
< 15 ms	93%	76%
< 20 ms	96%	84%
< 32 ms	99%	90%
< 64 ms	100%	95%

Table 5: Amount of corresponding segment boundaries that are lie within a defined time range. (a) manual segmentations (b) automatic segmentations

5. SUMMARY

Our approach of evaluating segmentation and transcription data of labelled speech has been shown. It is exemplified on the basis of manual and automatic (MAUS) transcriptions taken from the Phondat-II database of German read speech. The evaluation involves the definition of a reference. The degree of identity of chosen segment labels and the amount of deviation of segment boundaries respectively among manual transcriptions is the reference value to which the performance of the automatic segmentation system compared to manual transcriptions is related.

In a first part the evaluation refers to the degree of identity of transcription labels. It appeared that 88.4% of the automatic transcription of consonants is identical with manual transcriptions. The identity of manually chosen consonant labels comes to 94.8%, yielding a difference of 6.4%.

In a second part we looked at the degree of accuracy of segment boundaries which are identically labelled. Looking at segment boundaries in terms of transitions we get relevant information for the description of weaknesses of the automatic segmentation system which can than be tackled in a refinement-stage of the MAUS system.

The automatic segmentations of MAUS show 61% correct boundaries with an accuracy of 10 ms (human reference: 87%). An amount of 84% segment boundaries is correct within a range of 20 ms (human reference: 96%).

These are remarkable results obtained by an automatic segmentation system, which can be improved well-aimed through the specific information from the evaluation of segmentation and transcription data.

6. REFERENCES

1. Barry, W. J, Fourcin, A. J.: “Levels of Labelling”, Computer, Speech and Language 6, 1992, pp 1-14.
2. Eisen, B.: “Reliability of Speech Segmentation and Labelling at Different Levels of Transcription”, *Proceedings of EUROSPEECH 1991,* Berlin/Germany, pp. 673-676
3. Eisen, B., H. G. Tillmann, Chr. Draxler (1993): “Consistency of Judgements in Manual Labelling of Phonetic Segments: The Distinction between Clear and Unclear Cases”, *Proceedings of ICSLP 199*2, Banff/Canada, pp.871-874.
4. Kipp, A., M.-B. Wesenick, F. Schiel “Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora”, *these Proceedings of ICSLP 1996*, Philadelphia/USA.
5. Pompino-Marschall, B. (ed.): “PHONDAT. Verbundvorhaben zum Aufbau einer Sprachsignaldatenbank für gesprochenes Deutsch”, Forschungsbericht des IPSK (FIPKM) 30, 99 - 128, 1992.
6. Wesenick, M.-B., A. Kipp “Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals”, *these Proceedings of ICSLP 1996*, Philadelphia/USA.
7. Wesenick, M.-B., Schiel, F. (1994): “Applying Speech Verification to a Large Data Base of German to Obtain a Statistical Survey about Rules of Pronunciation”, *Proceedings of ICSLP 1994*, Yokohama/Japan, pp. 279 - 282.