The Quality of Multilingual Automatic Segmentation Using German MAUS

N. Beringer[1], *F. Schiel*[2]

[1] Institut für Phonetik und Sprachliche Kommunikation, [2] Bavarian Archive for Speech Signals (BAS), Schellingstr. 3, 80799 München, Germany

ABSTRACT

The goal of this work is to demonstrate the quality of multilingual automatic segmentations using the German MAUS system ([1, 5]) in order to substitute costly manually segmented data by automatically segmented corpora. In this study we investigated the influence of language specific HMMs in a cross-language task namely the automatic segmentations of English, French and Japanese with HMMs trained on German acoustic data. Given the orthographic transcription of an utterance we were able to produce quite good segmentations with the "wrong" acoustic models which will be described in detail in the following sections. The reason for this can either be based on the bigger influence of intra-/inter-speaker variability compared to the "interlingual variability" or on universal coarticulation processes as discussed below.

1. INTRODUCTION

By automatically segmenting according to phoneme various large corpora available from the BAS [4], we wanted to obtain on the one hand a reliable segmentation and labeling for applications in speech processing such as ASR and speech synthesis (e.g. PSOLA). On the other hand we want to expand our phonetic research also to other languages than German. We already showed in [6] and [9] that it is possible to get comparable results to manually segmented data in automatic segmentation for German. In these kinds of experiments we reach a correspondence with human labelers of approximately 78.5% by using the MAUS technique, while human inter-labeller agreement is about 80.4% ([1], [2]).

MAUS (Munich AUtomatic Segmentation) is an HMM-based system for the automatic segmentation of read or spontaneous speech. MAUS uses statistically weighted rewrite pronunciation rules for German and a Viterbi based alignment (HTK [3]) to automatically segment large speech corpora. The German version is trained on manually segmented data (approx. 1h40m of speech).

To segment non-German databases with MAUS, we usually have to train the system on the relevant acoustic models and to substitute the German rewrite pronunciation rules by a corresponding rule set, which is an expensive and time-consuming process.

The following section briefly describes the used data. Section 3 deals with the experiments we conducted in this investigation:

- the method
- the segmentation results for American English
- the segmentation results for French
- the segmentation results for Japanese

The results are interpreted in section 4 under the following aspects:

- intra- and inter-speaker variability vs. "interlingual variability"
- universal common coarticulation processes
- Influence of acoustic vs. pronunciation modeling

Finally, results and future work are discussed in the last section.

2. DATABASE

The German MAUS system ([1, 5]) was trained on approx. 30 h of unscripted speech namely on the German portion of the VERBMOBIL I corpus ([10]).

It is used to segment German spontaneous or read speech by using statistical rewrite pronunciation rules for German. The rules are trained on manually segmented data (approx. 1h40m of speech).

Since the German VERBMOBIL corpus contains more than 700 speakers the HMMs give a broad speaker-independent distribution over German acoustics.

For the experiments in this paper we use expert pronunciation rule sets without statistics for the language in question ([11]). The test data for English and Japanese are subsets of the corresponding portions of the VERBMOBIL II corpus (1 volume each). The French acoustic data recorded in the VERBMOBIL task as well as the pronunciation rules were produced during an exchange with the Institute of Phonetics, University Marc Bloch, Strasbourg, France. Since the latter are considerable smaller than the English and Japanese data, the results from French should be seen as preliminar.

3. EXPERIMENT

3.1 Method

Metze et al. have shown in ([12]) that it is possible to identify a language automatically by computing the confidence measure of the acoustic models of an utterance in automatic speech recognition. They also found out that using the German acoustic models can lead to some problems. English acoustic models seem to be very similar to German ones.

That led us to the following assumption:

- Given German acoustic models and pronunciation rules it should be possible to obtain quite good automatic segmentations for English based on a forced alignment.

To demonstrate our assumption we segmented not only the English data based on German acoustic models and pronunciation rules but also the French and Japanese data.

To evaluate the quality of the segmentation, we compared it to manual segmentations of the same corpus.

3.2 Results

It has already been shown elsewhere that a uniquely correct segmentation and labeling of an utterance does not exist because no two human experts are likely to produce exactly the same segmentation for the same utterance. Not even the same trained person will come to exactly the same transcription if asked to repeat the segmentation of the same utterance. In previous work it was shown that human labelers can reach a correspondence of about 80.4% ([13]).

Taking this value as reference it has already been shown that the German MAUS-System is statistically correct on manual segmentations in 97.5% of the segmentations for German with German acoustic models ([5]). Our results were quite encouraging as can be seen in table 1:

Language	correspondence in percent
French	86.02%
English	80.65%
Japanese	75.27%

Table 1: Comparison of automatic segmentation and manual segmentation of the same utterances compared to the mean correspondence of human labellers

The results in the French labelling task seem surprisingly high compared to the other languages. One reason for that may be the fact that the recorded speakers are from a region close to the German border around Strasbourg.

4. INTERPRETATION

The results of our experiment shows that acoustic features show considerable overlay in different languages.

There are – among others – two non-exclusive hypotheses to explain these results:

- Statistical: The intra-speaker and inter-speaker variability of acoustics and pronunciation is considerably higher than the 'interlingual variability'.
- Phonological: There exists a universal core of coarticulation processes common to all three languages that is reflected by the specific rule sets.

4.1 Statistically Based Hypothesis

To automatically segment utterances we usually use robust acoustic models which are trained on the result of a broad manual phonetic transcription. For example, HMMs do not consider either diacritics or fundamental frequency as caracteristic features. Comparing the standard SAM-PA phone systems the following table shows the number of equivalent phones. Note that because of the loss of diacritics the number of phones per system was about 45.

Languages	number of equivalent phones
ger-eng-jap-fra	17
ger-jap-fra	18
ger-eng-jap	18
ger-eng-fra	19
ger-jap	20
ger-fra	21
ger-eng	28

Table 2: Comparison of the number of equivalent phones in the language specific SAM-PA. ger stands for German, eng for English, jap for Japanese and fra for French.

Based on the fact that we use German acoustic models, we compared the phonetic alphabets of the languages in question with the German one.

It can be seen that about half of the equivalent phones of all languages are the same. Bilingual comparisons (respective to German) can have at most two third of equivalent forms as can be observed with German-English.

To illustrate the equivalent phonemes refer to table 3. It can be observed that we get a correspondence of plosives [b, d, g, p, t, k], nasals [m, n, N], fricatives [s, z, S], low and centralized vowels [a, e, @], the liquid [l] and the semivowel [j]:

Languages	equivalent strings
ger-eng-jap-fra	@, N, S, a, b, d, e, g, j, k, l, m,
	n, p, s, t, z
ger-jap-fra	@, E, N, S, a, b, d, e, g, j, k, l, m,
	n, p, s, t, z
ger-eng-jap	@, N, S, a, b, d, e, g, h, j, k, l, m,
	n, p, s, t, z
ger-eng-fra	@, N, S, a, b, d, e, f, g, j, k, l, m,
	n, p, s, t, v, z
ger-jap	@, C, E, N, S, a, b, d, e, g, h, j, k, l,
	m, n, p, s, t, z
ge-fra	@, E, N, O, S, a, b, d, e, f, g, j, k, l,
	m, n, p, s, t, v, z
ger-eng	@, I, N, S, U, a, aI, aU, b, d, e, f, g,
	h, i:, j, k, l, m, n, p, p:, r

Table 3: Equivalent phones in the language specific SAMPA.

Based on this results we can conclude that

- there is a low 'interlingual variability' and
- the high confusions of German and English in language identification found by [12] is probably caused by the high correspondence of the two phone systems, thus resulting in a high overlap of the acoustic models.

These findings suggest that the inter-lingual variability is low as we expected and especially low for the language pair German - English. To prove that this variability is mathematically lower than the variability within a language a direct comparison of intra-speaker and inter-speaker acoustic models would be necessary. Such a comparison is a difficult task and requires much more data from single speakers than was available for this investigation. However, it is well known that speaker-dependent ASR, as in dictation systems, fail if trained to the wrong speaker.

The models used in this experiment were trained to more than 700 speakers. Therefore we expect rather broad statistical distributions in the HMM states that might overlap an inter-lingual variability.

4.2 Phonologically Based Hypothesis

It has been shown elsewhere that gestures can be considerably reduced or even omitted if the sequence of gestures becomes too complex and if the new form does not lead to a homonym ([15]).

There are definitely universal sequences of gestures which are difficult per se, independent of the mother tongue. For example in German, especially in unprompted speech which is more relaxed than read speech, the wordfinal plosives may be left out, if the elision does not lead to an ambiguity in word semantics. This was shown in ([14]). Weak forms, i.e."in unstressed [function] words the distances the articulators travel are reduced to spaces closer to their neutral positions". Wordfinal consonants, especially those which need apical gestures, like "-t" in German "braucht" are often eliminated. This would not affect word perception. In other words after using the tongue dorsum as articulator an articulatory gesture produced by the tongue apex - a "more controlled and precisely tuned [...] therefore also more costly" gesture - is replaced by a "long oral closure of the dorsum". The preceeding fricative is not reduced because of its "acoustically and auditorily far more" distinctproperty which leads to a more salient value [p. 87f.].

Comparing the language specific rules there are in fact universal rules which can be seen in all languages, e.g. palatalization of plosives before high vowels, @-elision, centralization as well as deletion of (long) vowels in weak positions, assimilations of manner or place of articulation or the loss of plosives if occuring in wordend position. (For more details on the specific rule sets refer to [9], [11], [16]).

Of course, a universal core of coarticulation processes common to all observed languages is necessary for the observed results. But in principle there only have to be correspondences between the German rule set and the other language specific rule sets.

When comparing the language specific rules with the German rule set many similarities can be observed especially in English and French:

- Devoicing of plosives in wordend position. This includes both glottalization and elision of the corresponding phone.
- Fricativization of voiced plosives. This can be observed not only in English but also in German regional variants.
- Voicing of unvoiced consonants between vowels.
- Vocalization of r.
- Monophthongization of diphthongs. This phonetic process can be found in regional variants which are included in the German rule set.
- Centralization of vowels.

4. CONCLUSION

In the preceding sections we showed that it is possible to get quite satisfying automatic segmentations of several languages while using German acoustic models. We discussed two hypothesis for this phenomenon which are summed up as follows:

- There is a low 'interlingual variability' of the language specific phoneme sets
- The intra-speaker and inter-speaker variability is undoubtedly high but is compensated for by the robustness of the acoustic models.
- There exists an universal core of common coarticulation processes of all three languages which are caused by articulatory constraints.

- A correspondence of pronunciation rules of the German rule set to the other languages was observed as well.

We still have to investigate if the acoustic models themselves correspond as well. Therefore we need acoustic models of the individual speakers of all languages to compare their formant structures, fundamental frequencies etc. The average models for each language can then be compared across languages.

Furthermore, it would be interesting to quantify the contributions of acoustic vs. pronunciation modeling for the effect presented in this paper.

References

[1] Kipp, A. : Automatische Segmentierung und Etikettierung von Spontansprache; Shaker Verlag Aachen 1999.

[2] Beringer, N.; Schiel, F. (1999) Independent Automatic Segmentation of Speech by Pronunciation Modeling. Proc. of the ICPhS 1999. San Francisco. August 1999. pp. 1653-1656

[3] The HTK Book (for HTK Version 2.0), Cambridge University 1996

[4] F. Schiel (1998): Speech and Speech-Related Resources at BAS; Proceedings of the FIRST INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION 1998, Granada, Spain, pp. 343-349.

[5] F. Schiel (1999): Automatic Phonetic Transcription of Non-Prompted Speech. Proc. of the ICPhS 1999. San Francisco. August 1999. pp. 607-610

[6] Kipp, A., Wesenick, B., Schiel, F. (1997): Pronunciation Modeling Applied to Automatic Segmentation of Spontaneous Speech; in: Proceedings of the EUROSPEECH 1997, Rhodos, Greece, pp.1023-1026.

[7] M.-B. Wesenick, A. Kipp (1996): Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals; in: Proceedings of the ICSLP 1996. Philadelphia, pp. 129-132, Oct 1996.

[8] Beringer, N.; Schiel, F. ; Regel-Brietzmann P.: German Regional Variants - A Problem for Automatic Speech Recognition? in: Proceedings of the ICSLP 1998. Sydney, Dec 1998.

[9] Beringer, N.; Neff M.: Regional pronunciation variants for automatic segmentation; in Proceedings of the SECOND INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION 2000, Athens, Greece.

[10] German VERBMOBIL Corpus : www.phonetik.uni-muenchen.de/Bas

[11] Beringer, N; Neff M.; Ito T.: Generation of pronunciation rule sets for automatic segmentation of American English and Japanese to appear in: Proceedings of the ICSLP 2000. Beijing, Oct 2000.

[12] F. Metze, T. Kemp, T. Schaaf, T. Schultz, and H. Soltau: Confidence Measure based Language Identification in: Proceedings of the ICASSP 2000, Istanbul, June 5-9, 2000

[13] B. Eisen, H. G. Tillman, and C. Draxler: Consistency of judgements in manual labeling of phonetic segments: The distinction between clear and unclear cases, Proc of the ICSLP (Banff), 1992, pp. 871–874.

[14] Kohler, Klaus J.: Segmental Reduction in Connected Speech in German: Phonological Facts and Phonetic Explanations in: Hardcastle, Marchal: Speech production and Speech modelling, pp. 69-92 , 1990

[15] Michael Jessen: Die dorsalen Reibelaute [C] und [X] im Deutschen, in: Linguistische Berichte 117(1988), S371 - 396.

[16] M.-B. Wesenick (1996): Automatic Generation of German Pronunciation Variants; in: Proceedings of the ICSLP 1996. Philadelphia, pp. 125-128, Oct 1996.