HMMs for Automatic Phonetic Segmentation

Doroteo Torre Toledano[1]

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
200 Technology Sq. 02139 Cambridge, MA, USA
doroteo@sls.lcs.mit.edu

Luis A. Hernández Gómez

Grupo de Aplicaciones del Procesado de Señal
Dpto. Señales, Sistemas y Radiocomunicaciones
Universidad Politécnica de Madrid
Ciudad Universitaria s/n, 28040 Madrid, SPAIN
luis@gaps.ssr.upm.es

Abstract

This paper presents an analysis of the most frequently used approach in automatic phonetic segmentation – computing forced alignments using HMMs and features similar to those used in speech recognition. We start by analyzing the segmentation accuracy of context-dependent and context-independent HMMs, and proposing an explanation for the results. We focus our attention on the loss of correspondence between phones and context-dependent HMMs. This effect was already proposed to explain the surprisingly worse segmentation accuracy of context-dependent HMMs, given its clear superiority in speech recognition. We argue that this effect should lead to systematic segmentation errors. Therefore, we propose a new method, called Statistical Correction of Context Dependent Boundary Marks (SCCDBM), which partially corrects these systematic errors making segmentation results for context-dependent HMMs followed SCCDBM clearly superior to those obtained with context-independent HMMs. This observation empirically proves the existence of systematic segmentation errors and adds empirical evidence to the explanation for the worse segmentation accuracy of context-dependent HMMs. Finally, we analyze how speaker adaptation improves segmentation accuracy, and how speaker adaptation hardly modifies the systematic errors produced by context-dependent HMMs.

1. Introduction

Nowadays speech technology development strongly relies on corpus-based methodologies. In a speech corpus, almost as important as the speech itself is the information complementing it. In particular, phonetic segmentation and labeling are very useful because phones are usually considered the smallest constituents of speech. To date, the most precise way to obtain the phonetic segmentation and labeling is manually. However, manual phonetic labeling and (especially) segmentation are very costly and require much time and effort. For that reason, they have tried to be avoided.

In speech recognition Hidden Markov Models (HMMs) have avoided the need for manual phonetic segmentation. HMMs produce a segmentation, which although less precise than a manual segmentation, seems to be precise enough to train the HMMs because HMM training is an averaging process that tends to smooth the effects of segmentation errors (Cox et al., 1998).

For speech synthesis, however, the phonetic segmentation produced by HMMs is not precise enough. In the development of both concatenative acoustic unit inventories and prosodic models it is usual to select single examples instead of relying on an averaging process. As a consequence of the lack of averaging and error smoothing, a segmentation error may produce an audible error in the synthetic voice (Cox et al., 1998). This need for more precise segmentation has led speech synthesis to rely on manual segmentation for years. During the last few years, however, the need to develop new voices and languages quickly (Ljolje et al., 1997; Cox et al., 1998) and with the maximum quality (which frequently requires large inventories) has raised the interest in automatic segmentation techniques to partially automate the development of synthesis inventories and models.

2. Goal of the paper

This paper presents an analysis of the use of HMMs for automatic phonetic segmentation. It starts by reviewing the state of the art. Then, it analyzes and tries to explain the influence of the number of Gaussians per state and context-dependency on segmentation results. The theoretical explanation for the poorer segmentation accuracy of context-dependent HMMs is reviewed. We argue that this poorer accuracy is due to systematic errors that could be partially corrected with a new statistical correction method. We successfully test this possibility, achieving with context-dependent HMMs and this method higher segmentation accuracy than that achieved with context-independent HMMs. Finally, the role of speaker adaptation in automatic phonetic segmentation and how it interacts with this new correction technique are analyzed.

3. State of the art

The most frequent approach for automatic phonetic segmentation, and the one analyzed in this paper, is performing forced alignments of the speech and the orthographic transcription making use of phonetic HMMs and cepstral features (most frequently Mel-Frequency Cepstral Coefficients, MFCCs) very similar to those used in speech recognition (Ljolje et al., 1997; Cox et al., 1998; Angelini et al., 1993; Angelini et al., 1997).

There are two widely accepted adaptations of the HMMs and features to the particular problem of phonetic segmentation. One is using a higher frame rate than the one used in speech recognition to reduce quantization errors (Ljolje et al., 1997; Angelini et al., 1993; Angelini et al., 1997; Chou et al., 1998). The other is using context-independent HMMs instead of context-dependent HMMs because they provide better segmentation accuracy (Ljolje et al., 1997; Cox et al., 1998; Angelini et al., 1993; Angelini et al., 1997; Chou et al., 1998).

[1] Was with the Speech Technology Division of Telefónica Investigación y Desarrollo (R&D), c/ Emilio Vargas 6, 28043 Madrid, Spain.

It is worth mentioning that there is a significant number of authors that use completely different features or techniques (Farhat et al., 1993; Zue et al., 1989; Flammia et al. 1992, Karjalainen et al. 1998; Angelini et al., 1993; Malfrère et al. 1998). Others combine HMMs with other features or techniques (Boëffard et al., 1993; Malfrère et al., 1998; Chou et al., 1998).

Performance of segmentation algorithms can be assessed in an indirect way, for instance measuring the word error rate of a recognizer that uses the segmentation algorithm (Lee & Glass, 1998) or the subjective quality of a speech synthesizer generated making use of automatic segmentation (Cox et al., 1998; Boëffard et al., 1993). However, the most common and direct form of evaluation is comparing the segmentation to a manual segmentation to compute the segmentation accuracy as the percentage of boundaries with errors smaller than several values of tolerance (Ljolje et al., 1997; Angelini et al., 1993).

A comparison of the most commonly reported figure of merit (the percentage of boundaries with errors smaller than 20 ms.) in several research works reveals that best speaker-dependent results (around 96% of boundaries with errors below 20 ms.) have been achieved with HMMs (Angelini et al., 1997), synthesis and DTW (Angelini et al., 1993) and HMMs and refinement rules (Chou et al., 1998). For the speaker-independent case, best results (around 90% of boundaries with errors smaller than 20 ms.) have been obtained with HMMs (Angelini et al., 1993; Angelini et al., 1997).

4. Baseline system

In HMM-based phonetic segmentation it is common to use almost the same features, HMM topology and base phone set used for speech recognition. Most works use a higher frame rate than the one used in speech recognition to reduce quantization errors (Section 3). Our system computes a feature vector every 3 ms., using a 24 ms. Hamming window and a pre-emphasis coefficient of 0.97. The feature vector used in our system is the same used in our recognizer: 12 Mel-Frequency Cepstral Coefficients (MFCCs) with Cepstral Mean Normalization (CMN) and normalized log energy, as well as their first and second order differences, yielding a total of 39 components. For the HMM topology, our system uses 5 states, transitions from left to right and no skips. The high frame rate used allows for the use of 5 states without imposing important restrictions on phone duration. Output probability distributions in HMM states are modeled with mixtures of 1 to 6 diagonal covariance Gaussians. Finally, as most researches do, we rely on the same phone set used in our recognizer (24 Castilian Spanish phones). Our baseline system uses speaker-independent HMMs trained on carefully recorded speech (Section 9).

4.1. Results

Table 1 shows the segmentation accuracy (measured as the percentage of boundaries with errors smaller than several tolerance values in milliseconds) for context-independent and context-dependent HMMs with different numbers of Gaussians per state. These results were obtained on four corpora (*M1-80*, *M2-20*, *F1-20* and *M3Tot;* see Section 9 for corpora description). Results in bold face are the best results using different number of Gaussians for either context-dependent or context-independent HMMs. Shaded results are the best results obtained for a given number of Gaussians with either context-dependent or context-independent HMMs.

CD/CI	# Gauss	<5	<10	<20	<50	<100
Context-Independent	1	20.95	**44.84**	82.38	96.04	99.22
	2	**21.14**	44.68	**82.45**	96.39	99.38
	3	20.31	42.76	81.26	96.58	99.45
	4	19.17	41.77	80.56	96.62	99.45
	5	18.38	41.11	80.01	96.77	99.55
	6	18.12	40.38	79.41	**96.89**	**99.58**
Context-Dependent	1	26.61	**48.94**	77.89	97.45	99.52
	2	**26.68**	48.93	77.95	97.45	99.52
	3	26.26	48.42	**78.20**	**97.60**	99.50
	4	26.01	48.07	77.77	97.45	99.50
	5	25.86	48.03	77.61	97.53	99.55
	6	25.71	47.87	77.68	97.52	**99.56**

Table 1: Segmentation accuracy for context dependent and independent HMMs with different numbers of Gaussians.

With respect to the number of Gaussians per HMM state several trends can be observed for both context-dependent and context-independent HMMs. More Gaussians tend to produce better results when large tolerances (comparable to the duration of a phone) are considered, and worse results for small tolerances. For small tolerances, results steadily degrade as the number of Gaussians increase. This degradation is much faster for context-independent HMMs. For large tolerances, the tendency for improvement as the number of Gaussians increases is also stronger for context-independent HMMs.

With respect to context dependency, our experimental results show that context-independent HMMs only outperform context-dependent ones when the tolerance considered is around 20 ms. We have analyzed these results in more detail (analyzing results for 20 values of tolerance from 5 ms. to 100 ms., although the complete results are not presented here) finding three different tolerance zones where certain HMMs behave best. For small tolerances (5-10 ms.) context-dependent HMMs with fewer Gaussians behave best. For medium tolerances (15-30 ms.) context-independent HMMs with fewer Gaussians are better. Finally, for large tolerances (>35 ms.) context-dependent HMMs with more Gaussians tend to produce better results. This behavior, found in a global evaluation of phonetic segmentations using four speakers, was also found when results were analyzed speaker by speaker, with only small variations in the tolerance zones.

4.2. Factors influencing segmentation error

A possible explanation for the results obtained is based on the concurrency of different factors that are more or less important depending on the tolerance considered for the definition of the segmentation accuracy:

4.2.1. Coarticulation effects

Context-dependent HMMs can model coarticulation effects that cannot be modeled by context-independent HMMs, thus allowing a more detailed transition modeling that would explain the higher accuracy of context-dependent HMMs for small tolerances.

4.2.2. Loss of correspondence between phones and context-dependent HMMs

The fact that context-independent HMMs are preferred for speech segmentation (Section 3) is somewhat surprising, given that context-dependent HMMs model more accurately the spectrum dynamics and produce better results in speech recognition. This apparent paradox was explained theoretically in (Ljolje et al., 1997), where it was argued that the cause is the loss of correspondence, during the training process, between the context-dependent HMMs and the phones. Context-dependent HMMs are always trained with realizations of phones in the same context. For that reason, the HMMs don't have any clues to discriminate between the phone and its context. As a result the HMM (particularly the lateral states) can end up modeling part of other phones or not the whole phone. Context-independent HMMs, on the other hand, are trained with realizations of phones in different contexts. For that reason they should be able to discriminate between the phone to model (invariable in all training examples) and its context (which varies).

This would explain the better performance of context-independent HMMs in the intermediate range of tolerances (the range of tolerances most frequently reported in the literature).

4.2.3. Phone recognition accuracy

Large segmentation errors (comparable to the duration of a phone) are mostly due to phone misrecognition. Therefore, an increased phone recognition accuracy will lead to higher segmentation accuracy for large tolerances.

This would explain why context-dependent HMMs with more Gaussians tend to be better in the range of large tolerances and also why segmentation accuracy increases faster for context-independent than for context-dependent HMMs in that range – an increase in the number of Gaussians increases phone recognition accuracy more for context-independent HMMs than for context-dependent HMMs.

4.2.4. Spectral modeling near phone transitions

Results in the range of small tolerances tend to degrade as the number of Gaussians increase. One possible explanation for this observation may be that the spectrum near phone transitions is very variable and is better modeled with a reduced number of Gaussians. This would also explain why the degradation is faster for context-independent models, where the variation to model is larger due to coarticulation effects.

5. Statistical correction of context-dependent boundary marks (SCCDBM)

If the explanation for the decreased accuracy of context-dependent HMMs in phonetic segmentation (Section 4.2.2) is true, the root of this decreased accuracy is that context-dependent HMMs actually model the phone and part of its context or only part of the phone. This problem should lead to systematic segmentation errors. Therefore, it should be possible to statistically model the errors and use the resulting model to partially cancel them. This possibility gives rise to a new technique that we call *Statistical Correction of Context Dependent Boundary Marks* (SCCDBM). This technique comprises two steps: a training phase, where some statistical averages are estimated, and a boundary correction phase, where the phone boundaries produced by context-dependent HMMs are moved according to those estimated averages.

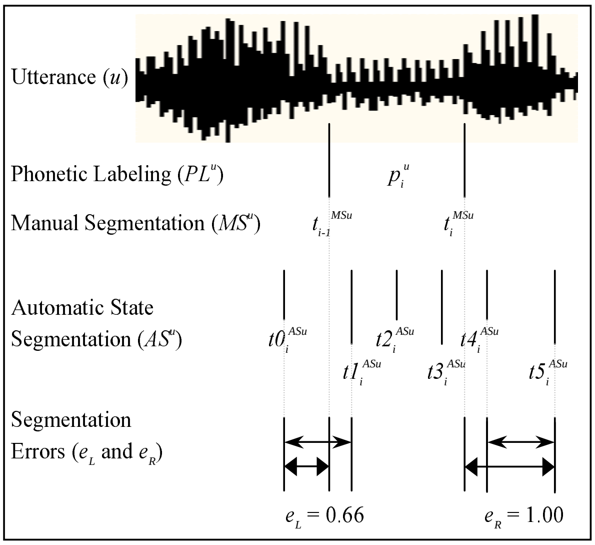

Figure 1: Statistical Correction of Context Dependent Boundary Marks (SCCDBM): Training phase.

Training (Figure 1) relies on a manually segmented corpus that is also automatically segmented at the state level using the context-dependent HMMs of interest.

Let's denote the set of training utterances as

$$U = \{ u_1, \dots, u_N \} \tag{1}$$

the phonetic labeling of an utterance u (which is considered known) as

$$PL^u = \{ p_1^u, \dots, p^u_{Pu} \} \tag{2}$$

the manual segmentation associated to an utterance u as

$$MS^u = \{ t_i^{Msu} \}_{i = 0, \dots, Pu} \tag{3}$$

and the automatic HMM-based segmentation of an utterance u at the state level (our HMMs have 5 states) as

$$\begin{aligned} AS^u = \{ & t0_i^{ASu}, t1_i^{ASu}, t2_i^{ASu}, t3_i^{ASu}, \\ & t4_i^{ASu}, t5_i^{Asu} \}_{i=1, \dots, Pu} \end{aligned} \tag{4}$$

where $t0_i^{ASu}$ and $t5_i^{Asu}$ correspond to the automatic boundaries for the phone p_i^u.

The first form of SCCDBM we tested computed the following segmentation errors for each phone p_i^u.

$$\begin{aligned} e_L(p_i^u) &= \max\left\{0, \min\left\{1, \frac{t_{i-1}^{MSu} - t0_i^{ASu}}{t1_i^{ASu} - t0_i^{ASu}}\right\}\right\} \\ e_R(p_i^u) &= \max\left\{0, \min\left\{1, \frac{t5_i^{ASu} - t_i^{MSu}}{t5_i^{ASu} - t4_i^{ASu}}\right\}\right\} \end{aligned} \tag{5}$$

Then, it computed the following averages for each of the context-dependent HMMs (determined by the left, p_L, central, p_C, and right, p_R, phones).

Method	<5	<10	<20	<50	<100
Context Indep. (CI)	21.75	49.48	83.85	96.58	99.46
Context Dep. (CD)	26.01	51.65	78.37	97.09	99.41
CD + SCCDBM	**41.32**	**67.81**	**88.90**	**98.67**	**99.80**

Table 2: Segmentation accuracy achieved with Statistical Correction of Context Dependent Boundary Marks (SCCDBM).

$$\hat{E}_L^{p_L,p_C,p_R} = \frac{1}{N_{p_L,p_C,p_R}} \sum_{\forall u \in U} \sum_{p_{i-1}^u = p_L, p_i^u = p_C, p_{i+1}^u = p_R} e_L(p_i^u)$$
$$\hat{E}_R^{p_L,p_C,p_R} = \frac{1}{N_{p_L,p_C,p_R}} \sum_{\forall u \in U} \sum_{p_{i-1}^u = p_L, p_i^u = p_C, p_{i+1}^u = p_R} e_R(p_i^u) \qquad (6)$$

where N_{p_L,p_C,p_R} is the number of training examples.

In the boundary correction phase, the automatic boundary between p_i^x and p_{i+1}^x in a test utterance *x different* from the ones used for training, is moved from its original position ($t5_i^{ASx} = t0_{i+1}^{ASx}$) to the position

$$t_i^{SCCDBMx} = t0_{i+1}^{ASx} + \hat{E}_L^{p_i^x,p_{i+1}^x,p_{i+2}^x} \cdot \left(t1_{i+1}^{ASx} - t0_{i+1}^{ASx}\right)$$
$$- \hat{E}_R^{p_{i-1}^x,p_i^x,p_{i+1}^x} \cdot \left(t5_i^{ASx} - t4_i^{ASx}\right) \qquad (7)$$

Although this first form of SCCDBM improved segmentation results, it required a huge amount of data to obtain reliable estimates. We later reduced the number of parameters to estimate and obtained increased correction capability making use of the state-sharing scheme used in our context-dependent HMMs. Instead of computing the averages in (6) for each triphone we computed similar averages for each different initial and final HMM state. This is the SCCDBM form for which results are presented in this paper.

Results have been obtained training the SCCDBM with one speaker (*M3Tot*) and applying that model to correct the segmentations produced by context-dependent HMMs on three other speakers (*M1-80*, *M2-20* and *F1-20;* see Section 9 for corpora description). Table 2 presents the segmentation accuracy (percentage of automatic boundaries with errors smaller than several tolerances) obtained with context-dependent HMMs, context-independent HMMs, and context-dependent HMMs followed by SCCDBM. Results presented were obtained with one Gaussian per state. We can observe that this technique substantially improves segmentation results for context-dependent HMMs. Moreover, observing the white bars in Figure 3, which represent the relative reduction of segmentation errors when SCCDBM is applied (see Section 7 for a more detailed description of Figure 3), apart from the 80% improvement found for very large tolerances (due to the correction of the majority of very few large errors), it can be seen that the most important improvements achieved with SCCDBM are in the intermediate range of tolerances, where we found the effect of the loss of correspondence between phones and context-dependent HMMs to be dominant in our experiments. All these results constitute empirical evidence that context-dependent HMMs produce systematic segmentation errors, and support the argument of the loss of correspondence between phones and context-dependent HMMs as the cause for the worse segmentation performance reported for this kind of HMMs. Results obtained with context-dependent HMMs followed by SCCDBM are better than those obtained with context-independent or with context-dependent HMMs alone *for all the tolerances* (Table 2). The same conclusions were found using other corpora for training and testing the SCCDBM, and also for other numbers of Gaussians per state. This result seems to indicate that SCCDBM can reduce so much the influence of the segmentation errors produced by the loss of correspondence between phones and context-dependent HMMs that they are no longer dominant for any range of tolerance. Our speaker-independent result for errors smaller than 20 ms. (88.90%) is very close to the best results found in the literature (around 90%). For one of the speakers analyzed we found 91.18% of segmentation errors smaller than 20ms.

Adaptation Technique	<5	<10	<20	<50	<100
No adaptation	43.43	71.63	91.18	97.58	99.65
MLLR 32	46.02	75.61	92.56	98.62	**100**
MLLR 64	46.71	75.78	93.08	98.96	**100**
MAP	44.64	73.70	92.39	98.62	**100**
MLLR 32 + MAP	**47.75**	**77.16**	**93.60**	99.13	**100**
MLLR 64 + MAP	47.06	76.82	**93.60**	**99.31**	**100**

Table 3: Segmentation accuracy achieved with different speaker adaptation techniques.

6. Speaker adaptation for speech segmentation

This section analyzes how two popular speaker adaptation techniques, Maximum Likelihood Linear Regression (MLLR) (Leggetter & Woodland, 1995) and Maximum A Posteriori (MAP) (Gauvain & Lee, 1994) behave in phonetic segmentation. Table 3 presents our experimental results for a female speaker (*F1-20*). These results were obtained with context-dependent HMMs with 3 Gaussians per state, and include SCCDBM trained for the speaker-independent HMMs on corpus *M3Tot*. Results show that MAP, MLLR and combinations of both speaker adaptation techniques improve segmentation accuracy for all tolerance values. Similar conclusions are reached with other corpora and other numbers of Gaussians per state. Figure 2 represents the relative reduction of segmentation errors achieved for the corpus *F1-20* with different speaker adaptation techniques. These results were obtained with 3 Gaussians per state and all include SCCDBM trained for the speaker independent HMMs on the corpus *M3Tot*. It can be observed that the most important improvements are in the range of large tolerances, where the dominant factor in the segmentation performance (Section 4.2) is the phone recognition accuracy. This is consistent with the proven ability of these speaker adaptation techniques to increase phone recognition accuracy of HMMs. We can also appreciate in Figure 2 that this very important improvement in the range of large tolerances is not accompanied by an important improvement in the range of small tolerances, perhaps

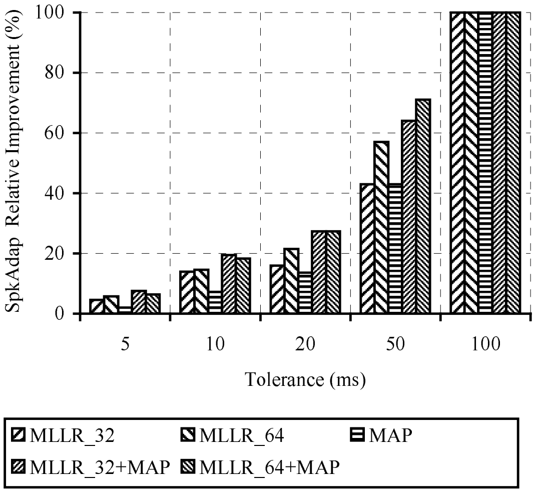

Figure 2: Relative reduction of segmentation errors achieved with different speaker adaptation techniques.

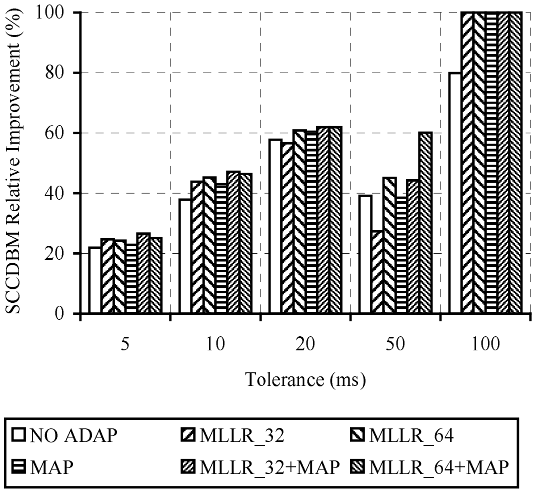

Figure 3: Relative reduction of segmentation errors achieved with SCCDBM trained for speaker-independent HMMs and applied to correct speaker-independent and speaker-adapted HMM-based segmentations.

because the more detailed modeling achieved with speaker adaptation is not very useful near the boundaries, where spectral variability is very high. Our speaker-dependent results for a tolerance of 20 ms. are reasonably good (93.6%) but still smaller than the best results found in the literature (around 96%).

7. Speaker adaptation and SCCDBM

To obtain both Table 3 and Figure 2 we have applied SCCDBM trained on the speaker-independent HMMs to correct the segmentations produced by speaker-adapted HMMs. This is based on the assumption that speaker adaptation doesn't affect substantially the systematic segmentation errors produced by the loss of correspondence between phones and context-dependent HMMs. A verification of this hypothesis can be seen in Figure 3, where the relative reduction of segmentation errors achieved by applying SCCDBM is compared when SCCDBM is applied to correct a segmentation produced by the same speaker-independent HMMs for which it was trained and when the same SCCDBM is applied to correct the segmentations produced by speaker-adapted versions of those HMMs. SCCDBM was trained on corpus *M3Tot* and results are computed on corpus *F1-20*. The HMMs used have 3 Gaussians per state. It can be observed that the improvements achieved are almost the same regardless of the use or not of these speaker adaptation methods. This confirms our assumption that the speaker-adaptation techniques used don't affect substantially the systematic errors produced due to the loss of correspondence between phones and context-dependent HMMs.

8. Conclusions

We have analyzed segmentation results for context-dependent and context-independent HMMs with different numbers of Gaussians per state and have proposed an explanation for the results based on the concurrency of several factors that are dominant for certain sizes of segmentation errors. We have addressed one of these factors, the loss of correspondence between phones and context-dependent HMMs (which explains the surprisingly poor segmentation results found for context-dependent HMMs), and have proposed and successfully tested a new technique, Statistical Correction of Context Dependent Boundary Marks (SCCDBM) that reduce the influence of this factor, making context-dependent HMMs plus SCCDBM clearly outperform context-independent HMMs for all error sizes considered. Then, we have analyzed speaker adaptation techniques, finding that they are able to correct very large segmentation errors but are not very effective to reduce small or intermediate segmentation errors. In particular, we have found that speaker adaptation techniques hardly influence segmentation errors that SCCDBM is able to correct.

Figure 4 shows the discrepancies found between two completely independent manual phonetic segmentations of corpus *M1-40-2* (which can be interpreted as the manual segmentation accuracy) and the segmentation accuracy of the HMM based methods analyzed in this paper: baseline (speaker-independent, context-dependent HMMs with 3 Gaussians per state), baseline plus SCCDBM, and baseline plus speaker adaptation (MLLR32+MAP) plus

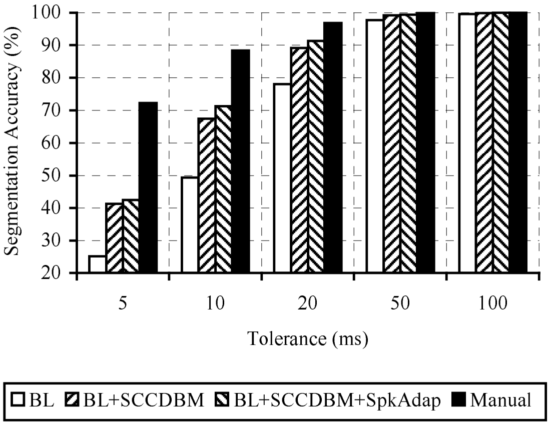

Figure 4: Comparison of manual and automatic segmentation accuracy for the baseline HMMs (BL), the baseline HMMs plus SCCDBM, and the baseline HMMs with speaker adaptation (SpkAdap) and SCCDBM.

SCCDBM. SCCDBM was trained on corpus *M3Tot* while results are presented for corpus *M1-40-2* (see Section 9 for corpora description). Improvements achieved over the baseline system by introducing SCCDBM and speaker adaptation are very important and lead to results close to the performance of a manual segmentation in the range of large tolerances. For small tolerances, however, results are still far away from manual accuracy. To improve segmentation accuracy for small tolerances, we have successfully used local refinement techniques (Toledano et al. 1998; Toledano, 2000; Toledano & Hernández, 2001) whose description is beyond the scope of this paper.

9. Corpora used

All the corpora used are in Castilian Spanish.

Speaker-independent HMMs were trained using a corpus containing 10 utterances and 75 isolated words for each of 1037 speakers, 619 male and 418 female. Speech was recorded with a microphone at 16 KHz in a clean environment and downsampled to 8 KHz.

Speaker adaptation experiments were carried out using four orthographically labeled mono-speaker corpora: *M1Tot* (389 sentences, 2531 words, 11090 phones, male speaker), *M2Tot* (454 sentences, 3089 words, 13490 phones, male speaker), *F1Tot* (532 sentences, 3685 words, 16199 phones, female speaker) and *M3Tot* (115 sentence sequences, 4086 words, 20175 phones, male speaker).

For segmentation evaluation some subsets of these mono-speaker corpora were phonetically labeled and segmented by hand: *M1-80* (80 sentences, 2464 phones, segmented by labeler *LabA*), several subsets of it as *M1-40-2* (40 sentences, 1221 phones, segmented by *LabA* and *LabB* to evaluate discrepancies in manual segmentations) and *M1-40-1* (40 sentences, 1242 phones, segmented by *LabA*), three sets that only differ in the speaker *M1-20* (a subset of *M1-40-1*), *M2-20*, *F1-20* (20 sentences, 599 phones, segmented by *LabA*), and *M3Tot*, manually segmented by a group of labelers.

10. Acknowledgements

This work was done at the Speech Technology Division of Telefónica Investigación y Desarrollo (R&D), which provided the corpora for the experiments. The authors would like to thank Luis Villarrubia for his advice on HMMs and Miguel Ángel Rodríguez and José Gregorio Escalada for their advice on speech synthesis.

11. References

Angelini B, Brugnara F, Falavigna D, Giuliani D, Gretter R and Omologo M, *Automatic Segmentation and Labelling of English and Italian Speech Databases*, In Proceedings EUROSPEECH 1993, pp 653-656.

Angelini B, Barolo C, Falavigna D, Omologo M and Sandri S, *Automatic Diphone Extraction for an Italian Text-To-Speech Synthesis System*, In Proceedings EUROSPEECH 1997, vol II, pp 581-584.

Boëffard O, Cherbonnel B, Emerard F and White S, *Automatic Segmentation and Quality Evaluation of Speech Unit Inventories for Concatenation-Based Multilingual PSOLA Text-To-Speech Systems*, In Proceedings EUROSPEECH 1993, pp 1449-1452.

Cox S, Brady R and Jackson P, *Techniques for Accurate Automatic Annotation of Speech Waveforms*, In Proceedings of the International Conference on Spoken Language Processing, 1998 Sydney (Australia), Vol V, pp 1947-1950.

Chou FC, Tseng CY and Lee LS. *Automatic Segmental and Prosodic Labeling of Mandarin Speech Database*, In Proceedings of the International Conference on Spoken Language Processing, 1998 Sydney (Australia), Vol IV, pp 1263-1266.

Farhat A, Pérennou G and André-Obrecht R, *A Segmental Approach Versus a Centisecond One for Automatic Phonetic Time-Alignment*, In Proceedings EUROSPEECH 1993, pp 657-660.

Flammia G, Dalsgaard P, Andersen O and Lindberg B, *Segment Based Variable Frame Rate Speech Analysis and Recognition Using a Spectral Variation Function*, In Proceedings of the International Conference on Spoken Language Processing 1992, pp 983-986.

Gauvain JL and Lee CH, *Maximum A Posteriori Estimation for Multivariate Gaussian Observations of Markov Chains,*, IEEE Transactions on Speech and Audio Processing, Vol. 2(2), pp. 291-298, 1994.

Karjalainen M, Altosaar T and Huttunen M, *An Efficient Labeling Tool for the Quicksig Speech Database*, In Proceedings of the International Conference on Spoken Language Processing, 1998 Sydney (Australia), Vol IV, pp 1535-1538.

Lee S and Glass J, *Real-Time Probabilistic Segmentation for Segment-Based Speech Recognition*, in Proceedings of the International Conference on Spoken Language Processing, 1998 Sydney (Australia), November 1998.

Leggetter CJ and Woodland PC, *Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models*, Computer, Speech and Language, 9(2), Apr. 1995, pp. 171-185

Ljolje A, Hirschberg J and Van Santen JPH, *Automatic Speech Segmentation for Concatenative Inventory Selection*, In Van Santen JPH et al. (eds), *Progress in Speech Synthesis*, Springer, 1997, pp 305-311.

Malfrère F, Deroo O and Dutoit T, *Phonetic Alignment: Speech Synthesis Based vs. Hybrid HMM/AN*, In Proceedings of the International Conference on Spoken Language Processing, 1998 Sydney (Australia), Vol IV, pp 1571-1574.

Toledano DT, Rodríguez MA and Escalada JG, *Trying to Mimic Human Segmentation of Speech Using HMM and Fuzzy Logic Post-Correction Rules*, In Proceedings of the 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, 1998 Jenolan Caves (Australia), pp. 207-212.

Toledano DT, *Neural Network Boundary Refining for Automatic Speech Segmentation*, In Proceedings of the International Conference on Acoustics Speech and Signal Processing 2000, Istanbul (Turkey), June 2000.

Toledano DT and Hernández LA, *Local Refinement of Phonetic Boundaries: A General Framework and its Application Using Different Transition Models*, In Proceedings EUSOSPEECH 2001, Aalborg (Denmark), September 2001.

Zue V, Glass J, Phillips M and Seneff S, *Acoustic Segmentation and Phonetic Classification in the SUMMIT System*, In Proceedings of the International Conference on Acoustics Speech and Signal Processing 1989, pp 389-392.