Comparison of Phonetic Segmentation Tools for European Portuguese

Luís Figueira and Luís C. Oliveira

L^{2}F Spoken Language Systems Lab.
INESC-ID/IST,
Rua Alves Redol 9, 1000-029 Lisbon, Portugal
{luisf,lco}@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt

Abstract. Currently, the majority of the text-to-speech synthesis systems that provide the most natural output are based on the selection and concatenation of variable size speech units chosen from an inventory of recordings. There are many different approaches to perform automatic speech segmentation. The most used are based on (Hidden Markov Models) HMM [1,2,3] or Artificial Neural Networks (ANN) [4], though Dynamic Time Warping (DTW) [3,4,5] based algorithms are also popular. Techniques involving speaker adaptation of acoustic models are usually more precise, but demand larger amounts of training data, which is not always available.

In this work we compare several phonetic segmentation tools, based in different technologies, and study the transition types where each segmentation tool achieves better results. To evaluate the segmentation tools we chose the criterion of the number of phonetic transitions (phone borders) with an error below 20ms when compared to the manual segmentation. This value is of common use in the literature [6] as a majorant of a phone error. Afterwards, we combine the individual segmentation tools, taking advantage of their differentiate behavior accordingly to the phonetic transition type. This approach improves the results obtained with any standalone tool used by itself. Since the goal of this work is the evaluation of fully automatic tools, we did not use any manual segmentation data to train models. The only manual information used during this study was the phonetic sequence.

The speech data was recorded by a professional male native European Portuguese speaker. The corpus contains 724 utterances, corresponding to 87 minutes of speech (including silences). It was manually segmented at the phonetic level by two expert phoneticians. It has a total of 45282 phones, with the following distribution by phonetic classes: vowels (45%), plosives (19.2%), fricatives (14.6%), liquids (9.9%), nasal consonants (5.7%) and silences (5.5%). The data was split in 5 training/test sets — with a ratio of 4/1 of the available data, without superposition. For this work we selected the following phonetic segmentation tools:

Multiple Acoustic Features–Dynamic Time Warping (MAF–DTW)**:** tool that improves the performance of the traditional DTW alignment algorithm by using a combination of multiple acoustic

A. Teixeira et al. (Eds.): PROPOR 2008, LNAI 5190, pp. 252–255, 2008.

features depending on the phonetic class of the segments being aligned [5]. The implementation of the MAF–DTW used in this experiment uses a synthetic European Portuguese male voice from a different speaker than the recorded in the corpus;

Audimus: is a speech recognition engine that uses a hybrid HMM/Multi-Layer Perceptron (MLP) acoustic model combining posterior phone probabilities generated by several MLP's trained on distinct input features [7,8]. The MLP network weights were re–estimated to adapt the models to the speaker;

Hidden Markov Model Toolkit: (HTK) [9], using unsupervised speaker-adapted, context-independent Hidden Markov Models (HMM). The models were adapted based on initial segmentations generated by the MAF–DTW tool. The models have ergodical left–right topology, with 5 states each (3 emitting states);

eHMM: phonetic alignment tool oriented for speech synthesis tasks [10] , developed in Carnegie Mellon University and distributed together with a set tools for building voices for Festival, called Festvox 2.1 [11]. The adopted model topology is the same as described for HTK; eHMM was also used doing acoustic model adaptation to the speaker.

In Table 1 we present the overall performance of each segmentation tool. From this table, it can be seen that the MAF–DTW is the tool with the worst performance in terms of Absolute Mean Error (AME): 41ms. This value is almost twice as much as the second worst result (eHMM). This was already expected, as DTW algorithms are usually very accurate, but simultaneously prone to gross labelling errors, when compared to speaker adapted algorithms [3]. Audimus has the best AME results, and also the smaller standard deviation results, showing that its errors are not widely spread (unlike DTW's). Both HMM based segmentation tools (eHMM and HTK) have a similar behavior.

Each tool's perfomance was evaluated for all the transition types. This study allowed the creation of a new segmentation tool by choosing the best tool for each transition type — using the highest number of borders inside the 20ms tolerance to the manual segmentations as the criterion. Table 2 shows the configuration of this segmenter (S1). Its overall results show that though its AME (16.60ms) is worst than Audimus' or eHMM's, there is an improvement in the number borders placed inside the 20ms error threshold (82.5%). This is due to the fact that the criterion used to choose the best segmenter for each transition is the 20ms error threshold performance, and not the AME. The S1 segmenter's composition shows

Table 1. Absolute Mean Error (AME), Root Mean Square Error (RMSE), Standard Deviation (σ) and borders with error below the 20ms tolerance ($< 20ms$)

	AME(ms)	**RMSE**(ms)	σ(ms)	$< 20ms$(%)
DTW	41.18	117.23	109.76	64.1
eHMM	20.54	33.07	25.92	68.1
HTK	15.44	24.00	48.9	76.9
Audimus	15.23	22.48	16.54	75.9

Table 2. S1 configuration: best combination of segmentation tools

	Nasal	**Fricative**	**Liquid**	**Plosive**	**Vowel**	**Silence**
Nasal	HTK	eHMM	eHMM	HTK	HTK	HTK
Fricative	Aud	Aud	eHMM	DTW	HTK	DTW
Liquid	Aud	eHMM	eHMM	Aud	HTK	Aud
Plosive	eHMM	HTK	HTK	HTK	HTK	HTK
Vowel	Aud	eHMM	Aud	Aud	Aud	DTW
Silence	eHMM	eHMM	eHMM	Aud	DTW	—

Table 3. SoM2 configuration: best combination of simple/pairs of segmentation tools

	Nasal	**Fricative**	**Liquid**	**Plosive**	**Vowel**	**Silence**
Nas	HTK	eHMM	eHMM	eHMM, HTK	Aud, HTK	eHMM,HTK
Fri	Aud	Aud	eHMM, Aud	DTW, Aud	HTK	DTW
Liq	Aud, HTK	eHMM, Aud	eHMM	Aud, HTK	Aud, HTK	Aud
Plo	eHMM	eHMM, Aud	Aud, HTK	eHMM, Aud	Aud, HTK	HTK
Vow	Aud, HTK	eHMM, Aud	Audi , HTK	Aud, HTK	Aud, HTK	DTW, HTK
Sil	eHMM	eHMM	eHMM	DTW, Aud	DTW, HTK	eHMM

that, as expected, the tools that involve acoustic model training have a better performance, though the DTW based algorithm performed better in some phonetic transitions — namely Fricative–Plosive, Silence–Vowel, Vowel–Silence and Fricative–Silence. The most important conclusion was that no segmentation tool obtained far superior results than the others: every tool had some transitions in which it performed better than any of the others, and transitions in which it performed worse.

Another configuration we studied was which pairs of segmenters obtained better results when its borders were combined linearly—*i.e.* for each transition the border was placed in the the average value of the two segmenters which yielded better results — again the criterion being the number of border inside the 20ms threshold. This new segmenter (M2) obtains better results than any of the individual segmenters, and even better than S1's, with an AME of 13.95ms, and 84.3% of the phonetic transitions with an error below 20ms.

The final configuration studied was the best combination of a single tool or the average of a pair of tools (SoM2). This presented the best results on the number of borders placed correctly: 84.6%. Its AME is 14.3ms, which is only worse when compared to the M2 configuration; Tab. 3 shows the configuration of SoM2.

In the future we plan to expand this work to more databases, to ensure its validity for different speakers of both genders. We also plan to use this method in larger speech inventories, so that we are able to measure its effect on the output speech quality. Another research topic will be using a combination of multiple individual segmentation tools to evaluate the confidence of third–party segmentations of speech databases.

Keywords: Automatic Phonetic Segmentation, Speech Synthesis, Hidden Markov Models, Dynamic Time Warping.

Acknowledgments. The authors would like to thank M. Céu Viana and Helena Moniz for providing the manually aligned reference corpus. The authors would also like to thank Hugo Meinedo and Sérgio Paulo for providing some of the tools used in this study. This work was funded by PRIME National Project TECNOVOZ number 03/165.

References

1. Toledano, D.T., Gómez, L.A., Grande, L.V.: Automatic phonetic segmentation. IEEE Transactions on Speech and Audio Processing 11 (November 2003)
2. Huggins-Daines, D., Rudnicky, A.I.: A Constrained Baum-Welch Algorithm for Improved and Efficient Training. In: Proc. Interspeech 2006s-9th International Conference on Spoken Language Processing, Pittsburgh, USA (2006)
3. Black, A.W., Kominek, J., Bennett, C.: Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis. In: Proc. Eurospeech, Geneva, Switzerland, pp. 313–316 (2003)
4. Malfrre, F., Deroo, O., Dutoit, T.: Phonetic alignment: speech synthesis based vs. hybrid HMM/ANN. In: Proc. 5th International Conference on Spoken Language Processing (1998)
5. Paulo, S., Oliveira, L.C.: DTW-based Phonetic Alignment Using Multiple Acoustic Features. In: Proc. Eurospeech, Geneva, Switzerland, pp. 309–312 (2003)
6. Adell, J., Bonafonte, A.: Toward Phone Segmentation for Concatenative Speech Synthesis. In: Proc. 5th ISCA Workshop on Speech Synthesis (2004)
7. Neto, J.P., Martins, C., Meinedo, H., Almeida, L.B.: AUDIMUS — Sistema de Reconhecimento de Fala Contínua para o Português Europeu. In: PROPOR 1999 - IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada, Évora (1999)
8. Meinedo, H., Caseiro, D., Neto, J.P., Trancoso, I.: AUDIMUS.Media: A Broadcast News Speech Recognition System for the European Portuguese Language. In: Mamede, N.J., Baptista, J., Trancoso, I., Nunes, das Graças Volpe Nunes, M. (eds.) PROPOR 2003. LNCS, vol. 2721, pp. 9–17. Springer, Heidelberg (2003)
9. Young, S., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3.2). Cambridge University Engineering Department (2002)
10. Prahallad, K., Black, A.W., Ravishankar, M.: Sub-phonetic Modeling for Capturing Pronunciation Variations for Conversational Speech Synthesis. In: Proc. ICASSP (2006)
11. Black, A.W., Lenzo, K.A.: Building Synthetic Voices, For FestVox, 2.1 edn. Language Technologies Institute, Carnegie Mellon University and Cepstral, LLC (2006), `http://www.festvox.org`