Repairing Tracheoesophageal Speech Duration

Arantza del Pozo and Steve Young

Cambridge University Engineering Department
Trumpington Street, Cambridge, England, CB2 1PZ
{ad371; sjy}@eng.cam.ac.uk

Abstract

This paper describes an investigation into the repair of the prosodic limitations of tracheoesophageal (TE) speech. The proposed repair algorithm modifies TE phone durations based on the predictions of regression trees built from non-pathological data. Acoustic and language modelling refinements for improved TE phone recognition, studies of feature relevance for duration prediction and a robust duration modification method are also presented. Objective and subjective evaluation of results show that the duration pattern of the repaired sentences is closer to normal and perceptually preferred to the original in terms of overall rhythmic naturalness.

1. Introduction

Tracheoesophageal (TE) speech is the most frequently used voice restoration technique after total laryngectomy. It involves the insertion of a voice prosthesis, which enables the use of pulmonary air for voice production as in laryngeal speech. Despite often being cited as the alaryngeal speech alternative most comparable to normal, its quality and intelligibility are still significantly lower than those of laryngeal speech [1].

Because of the inability to properly control the pharingoesophageal segment acting as the new voice source after the removal of the vocal cords, excitation related features have been thought to be the main cause of degradation and previous TE speech repair attempts [2] have mainly focused on this limitation.

However, prosodic deviations also contribute to the reduction in quality. Despite being able to use pulmonary air for speech production, the need to control the voice prosthesis in order to switch between speaking and breathing affects the duration pattern of the resulting TE speech. In general, TE speakers tend to stop more often, produce vowels with longer duration, speak with slower rates than normal subjects and sometimes rush the last phones before phrase breaks.

As far as we know, there have been no previous attempts to repair the TE speech duration pattern. In this paper, we present a novel algorithm capable of automatically repairing the duration deviations of TE speech and producing utterances which have been found to be preferred in terms of rhythmic naturalness.

The structure of the paper is as follows. Section 2 describes the speech corpus and preliminary experiments which motivated the duration repair approach. An investigation into TE phone recognition, the adopted duration modelling technique and a method to cope with recognition errors are presented in Sections 3, 4 and 5 respectively. The duration repair algorithm is then evaluated in Section 6 and finally, conclusions are presented in Section 7.

2. Data and preliminary experiments

2.1. TE speech corpus

A parallel corpus of 11 normal and 13 TE speakers reading the Rainbow Passage and a small set of descriptive sentences was used throughout the experiments presented in this paper. The 28 sentences available per speaker were divided into training (23 utterances) and test (5 utterances) sets of normal and TE speech, which were used to build, adapt and test the various duration models and speech recognition systems.

2.2. Preliminary experiments

Measures of TE speech duration have shown that the main differences compared to normal are shorter maximum phonation times, longer vowel durations and slower rates [3]. In addition, they generally pause longer and more often and sometimes rush the last phones before breaks.

Possible approaches to repair such deviations are (a) to derive a set of rules to modify the duration features found to be abnormal (i.e. reduce vowel durations and pauses, increase speech rate and durations of phones before breaks) or (b) to substitute TE phone durations with their corresponding normal values.

The difficulty within the rule-based method lies in obtaining adequate reduction/increase rules and ratios, which generally differ per speaker and per sentence. Experiments with this approach resulted in unnatural duration contours which despite normalising the deviant duration features, nevertheless ruined sentence rhythm.

Transplantation of average normal phone duration contours obtained from the parallel corpus achieved better results. Informal listening of transplanted utterances showed an overall improvement, which increased the naturalness of the original TE samples. This method not only coped with the observed TE durational problems but also preserved the rhythmic structure of the sentences.

The proposed TE speech duration repair algorithm is an attempt to automate the preliminary transplantation experiment. However, this method presumes that TE phone segmentations and normal phone durations are known. Hence, to use the method in a real-time repair application where the transcription of the input speech is unknown, the TE phones need to be recognised and their normal durations need to be predicted. In addition, methods are needed to provide robustness to recognition and prediction errors. The adopted recognition, duration prediction and robust modification techniques are described in the next sections.

3. Tracheoesophageal phone recognition

Previous work on automatic TE speech recognition by Haderlein et al. [4] involved adapting a speech recogniser trained on normal speech to single TE speakers by unsupervised HMM interpolation. They obtained poor results in terms of word accuracy, with an average value of 36.4 %. Our preliminary word level recognition tests on TE speech also showed that extracting usable orthographic transcriptions was not feasible. Hence, the focus here is on obtaining the best possible TE phone recognition.

We have explored various systems and techniques in order to achieve best results. Our baseline recogniser is a monophone system trained on the WSJCAM0 corpus [5] of normal non-pathological speech. In addition, normalization and adaptation techniques and several acoustic and language models have been tested. In order to measure and compare performance of the different systems, two new metrics which not only measure recognition and segmentation accuracy but also take duration prediction errors into account have been used.

SYSTEM	FEATURES
BL	Baseline monophone HMM
R1	BL+CMN+CMLLR
R2	R1+bigram LM
R3	R1+trigram LM
R4	triphone HMMs+CMLLR+word trigram LM

Table 1: *The recognisers tested and their corresponding features*

3.1. Measuring performance

Performance of speech recognisers is generally measured by comparing the output string with a manually transcribed reference and counting the percentage of correctly recognised, substituted, inserted and deleted labels. These measures only take recognition of the correct labels into account, ignoring segmentation accuracy or the implications of errors in a duration modification task.

Automatically derived transcriptions can also be regarded as consisting of a set of correctly recognised and segmented sections with error segments in between, in which phones have been wrongly segmented and/or misrecognised. Differences between the durations predicted within these error segments and their correct counterparts are the cause of the perceptual artifacts produced when inaccurate phone label transcriptions are used instead of force-aligned segmentations based on accurate reference transcription.

For our speech repair application, the recogniser not only needs to recognise the correct phones, but also accurately detect their boundaries. In addition, it should try to minimise the duration prediction differences within the error segments. We have used the following two measures which take these requirements into account to evaluate and compare the different recognisers:

- **Segmentation and Prediction Correctness (SPC):** measures the percentage number of phones which have been correctly recognised, with segmentation boundaries lying within a threshold distance of the reference values. An ideal recogniser would correctly recognise and segment all phones, and thus have an SPC=1.

$$SPC = \frac{\sum_{i=1}^{NP} r(i) \cdot s(i)}{NP} \quad (1)$$

where r and s are boolean variables equal to one if a particular phone has been correctly recognised or segmented respectively, and NP is the total number of phones in a sentence.

- **Segmentation and Prediction Error (SPE):** sums the differences in duration prediction of the error segments with respect to the reference values throughout the utterance and normalises its value by the total number of phones in the sentence.

$$SPE = \frac{\sum_{s=1}^{ES} \left| D(s) - \sum_{i=1}^{ESP} d(i) \right|}{NP} \quad (2)$$

where ES is the number of error segments in the sentence, ESP is the number of recognised phones in a particular error segment, D is the duration predicted by the reference transcription in a segment and d is the duration prediction of a recognised phone.

The best recogniser will be the one which achieves maximum SPC and minimum SPE values.

3.2. System comparison

In order to improve recognition of the baseline monophone system, the following techniques were explored. Firstly, cepstral mean normalisation (CMN) was used to normalise the recording conditions of the training and test data. Also, as in [4], adaptation techniques were applied to compensate for the acoustic differences between normal and TE speech. The TE training data was used to adapt the baseline models to each of the TE speakers. Due to the limited amount of data (just 23 sentences per speaker) available for adaptation, constrained maximum likelihood linear regression (CMLLR) was used. Linear transformations were obtained after 3 iterations, each using two regression classes for phones and silences/short-pauses. These give the R1 system listed in Table 1.

Secondly, phone level bigram and trigram language models (LM) trained on the WSJCAM0 corpus were introduced to give R2 and R3 in Table 1, respectively.

Finally, a system based on triphone HMMs and the word trigram LM described in [6] was tested. These triphone models were trained on a very large corpus and then adapted to each TE speaker using CMLLR. As well as being better trained than in the other systems, the use of a word level LM has the potential to provide better phonotactic constraints than a phone level LM. This final system is R4 in Table 1.

	BL	R1	R2	R3	R4
SPC [%]	0.1634	0.3129	0.3249	0.3340	0.5148
SPE [ms]	39.444	29.713	27.329	26.682	14.257

Table 2: *Evaluation of recogniser performance*

The overall recognition performance of the different systems on the TE test set was compared. As shown in Table 2, normalization and adaptation almost doubled the baseline performance. The addition of bigram and trigram LMs further improved the SPC and SPE results. However, R4 achieved the

biggest improvement and best overall performance. These results show the value of using refined acoustic and language models to compensate for the small amount of TE data available for adaptation.

4. Duration Prediction

Modelling and predicting duration is a difficult task, since phone segmental durations are extremely dependent on many factors such as context, positional features and stress. Duration models have mainly been applied to predict timing in text-to-speech (TTS) systems. Early TTS implementations mostly employed rule-based duration models. Despite their reasonable performance, such models often over-generalize and cannot handle exceptions without becoming too complicated. For these reasons, computational progress and availability of large speech corpora has favoured the development and increased use of data driven approaches in state-of-the-art applications. Among these, we have chosen to use classification and regression trees (CART) [7] to predict phone durations in our repair system for two main reasons: because standard tools for their generation exist and because the derived trees can be interpreted and used to determine the most relevant features.

TTS CART duration models consider features which can be extracted from text. Unfortunately, as noted earlier, the high word error rates of the TE word level transcriptions prohibit their use in a practical duration repair system. As a result, only recognizable phone level information such as phone identity, identities of the previous and next phones and position of the phones in the sentence is available. In addition to these contextual and positional factors, the use of pitch and rms energy features has also been explored, in an attempt to incorporate some kind of stress information.

Different combinations of the available features were used to build five regression trees (T1, T2, T3, T4 and T5) and investigate phone level feature relevance for duration prediction. The trees were built using the Matlab implementation of CART [8]. Short pauses (SP) were not regarded as phones and were modelled independently in a parallel tree TSP. Table 3 provides a more detailed description of the different tree features. The normal speakers' training set was used as training data. Phone segmentation was achieved by force-aligning each sentence with the baseline recogniser BL. Speaker adapted versions of this model were used for the segmentation of TE speech.

TREE	FEATURES	
T1	F1	phone identity
T2	F2	F1 + prev and next phone identities (converted to broad classes)
T3	F3	F2+ position of phone in sentence (first 5 phones / last 5 phones / rest)
T4	F4	F3+pitch (positive slope / no slope / negative slope)
T5	F5	F4+energy (positive slope / no slope / negative slope)
TSP	FS	num of phones since prev sp, num of phones until next sp

Table 3: *Description of trees and corresponding features*

Tree performance was evaluated against the TE test set, computing the average mean squared error (MSE) between the mean normal durations used for transplantation and the predicted values. Results showed that T3, followed by T2 and T1, predicted phone durations closest to the transplanted values, revealing that phone context and positional information improve duration prediction (see Table 4). However, differences between them were not large and phone identity appeared to be the most relevant feature in the three cases. On the other hand, the addition of pitch and energy features decreased performance, showing that linear regression of phone pitch and intensity contours does not appropriately model lexical stress as we were hoping.

	T1	T2	T3	T4	T5
MSE [ms]	0.788	0.695	0.570	2.174	1.535

Table 4: *Evaluation of tree duration prediction*

TE sentences whose force-aligned phone durations were substituted by those predicted by T3 and TSP were informally found to be perceptually indistinguishable from their corresponding transplanted versions, demonstrating the validity of the adopted duration prediction approach. However, even when the best recognised segmentations from R4 were used instead of the force-aligned labels, phone recognition errors caused durational artifacts, emphasizing the need for a robust modification method capable of taking recognition errors into account.

5. Robust duration modification

One way to reduce the duration artifacts is to incorporate phone recognition confidence information in the repair process, and to modify durations accordingly. Such a method can be described by the following equation

$$d_N = \alpha \cdot d_P + (1-\alpha) \cdot d_O \quad (3)$$

where d_N is the new duration, d_P is the predicted duration, d_O is the original duration and α is a confidence measure.

The main difficulty with this technique lies in obtaining appropriate values of α. TE phone duration probability distributions and confidence scores can be used to compute the confidence measure. In addition, information on phone confusions can also be incorporated from phone level confusion networks. An analysis of the correlation between recognition errors and these features revealed that high confidence scores and duration probabilities corresponded to correctly segmented and labelled phones, while low duration probabilities or low confidence scores generally coincided with insertions, deletions and substitutions. Also, the correct phone was often included in the phone confusion lists. As a result, α was computed as the mean of the duration probability and confidence score for each phone and d_P was calculated as the average of the durations predicted by the confused phones.

The described robust modification (RM) technique was used to modify the durations of the phones recognised by R3 and R4. These systems will be referred to as RM1 and RM2 respectively for comparison purposes (see Table 5).

SYSTEM	FEATURES
RM1	R3+RM
RM2	R4+RM

Table 5: *Robust modification systems*

The application of this technique very considerably reduced durational artifacts. Even though converted

utterances did not perceptually match the transplanted versions, informal listening showed that the main TE duration deviations described in Section 1 were mostly repaired without additional artifacts, resulting in more natural duration contours overall. Also, sentences modified with RM1 and RM2 were found to be perceptually almost indistinguishable.

6. Evaluation

6.1. Objective evaluation

In order to test the performance of the different repair systems, the MSE between the repaired phone durations and the transplantation values was computed. Results in Table 6 show that the proposed repair technique reduces the MSE overall, bringing TE duration contours closer to those of the average normal speaker. In addition, the application of robust duration modification further improves results.

SYSTEM	MSE [ms]
original TE speech	10.080
R3	5.873
R4	3.913
RM1	3.994
RM2	3.186

Table 6: *Evaluation of repair systems*

6.2. Subjective evaluation

A perceptual test using a panel of 20 listeners was carried out in order to perceptually evaluate our repair algorithm. It consisted of two parts. In the first, randomly ordered triplets of original (O), transplanted (T) and repaired (R) versions of the same utterance were presented. Instances of sentences transformed using RM1 and RM2 were randomly selected for the repaired category. The subjects ranked each sentence from 1 to 5 in terms of the naturalness of their rhythm, 1 being very natural and 5 very unnatural. The aim of this part was to find out which duration repair approach naïve listeners preferred and if the application of RM1 or RM2 made a significant difference perceptually. Tables 7 and 8 show the results obtained overall.

	RANKING SCORE
T	2.65
R	3.15
O	3.68

Table 7: *Ranking scores*

	>	=	<
T - R	**0.54**	0.22	0.24
R - O	**0.52**	0.31	0.17
T - O	**0.66**	0.20	0.14

Table 8: *Preference test results*

As expected, T sentences were found to have the most natural rhythm followed by the R versions, both beating the O duration patterns. T utterances were preferred 66% and 54% of the time over the corresponding O and R ones, while R versions were ranked higher than the O in 52% of the choices. No significant differences were observed between the two repair methods RM1 and RM2.

In the second part of the perceptual test, listeners had to choose specifically between pairs of utterances repaired with RM1 and RM2. Subjects found it very hard to distinguish between systems and respectively preferred RM1 and RM2 48% and 52% of the time. These results and the lack of preference correlation found in the previous part show that, despite using more elaborated acoustic and language models, R4 achieves the same perceptual results as R3 when robust modification is adopted to reduce the impact of recognition errors.

7. Conclusions

In this paper a method to automatically repair TE speech durations has been presented. The basic idea is to substitute deviant TE phone durations with those predicted by a CART tree constructed from normal data. The real-time requirement of the speech repair framework brings the issue of recognition errors into play and prevents the use of text-based features employed in TTS duration models. Solutions to these problems have been proposed. Evaluation of the proposed repair algorithm has indicated a preference for the converted duration patterns. However, there is still room for improvement in the recognition step. If sufficiently accurate word level transcriptions could be obtained, text-based duration features could be included in the decision trees and this could yield further improvements.

8. Acknowledgements

This work was supported by a researcher training grant from the Government of the Basque Country. The authors thank the volunteers of the perceptual tests for their assistance.

9. References

[1] C.J. Van As, “Tracheoesophageal speech: A multidimensional assessment of voice quality”, PhD thesis, University of Amsterdam, 2001

[2] A. del Pozo and S. Young, “Continuous Tracheoesophageal Speech Repair”, EUSIPCO 2006

[3] J. Robbins, H.B. Fisher, E.C. Blom and M.I. Singer, “A comparative acoustic study of normal, esophageal and tracheoesophageal speech production”, Journal of Speech and Hearing Disorders, vol. 49, pp. 202-210, 1984

[4] T. Haderlein, S. Steidl, E. Nöth, F. Rosanowski and M. Schuster, “Automatic Recognition and Evaluation of Tracheoesophageal Speech”, In Text, Speech and Dialogue (TSD 2004), Proceedings LNAI 3206, Springer, Berlin, Heidelberg, pp. 331-338, 2004

[5] WSJCAM0 corpus, http://svr-www.eng.cam.ac.uk/~ajr/wsjcam0/wsjcam0.html

[6] D.Y. Kim, H.Y. Chan, G. Evermann, M.J.F. Gales, D. Mrva, K.C. Sim and P.C. Woodland, “Development of the CU-HTK 2004 Broadcast News Transcription System”, ICASSP 2005

[7] M.D. Riley, “Tree-based modeling for speech synthesis”, In G. Bailly, C. Beno it, and T. Sawallis (Eds.), Talking machines: Theories, models and designs, pp. 265-273, 1992

[8] Matlab Statistical Toolbox, http://www.mathworks.com