Proceedings of the Institute of Acoustics

EVALUATION AND OPTIMISATION OF A SEGMENTER FOR A PC-BASED PRONUNCIATION TEACHING SYSTEM

F R McInnes (1), F Carraro (2), S M Hiller (1) & E J Rooney (1)

(1) Centre for Speech Technology Research, University of Edinburgh, UK
(2) Alcatel FACE, Salerno, Italy

1. INTRODUCTION

In the SPELL system for foreign language pronunciation teaching [1], the student's pronunciation of each phrase is evaluated by locating salient phonetic segments in the utterance and measuring the values of various acoustic parameters in and around these segments. To make this possible, a segmenter is incorporated, which finds the best-matching phonetic transcription of the phrase together with the corresponding division of the utterance into segments.

This paper presents the results of a series of experiments in which the performance of this segmenter was evaluated both in an intra-language configuration (using English language data from native English speakers) and in an inter-language configuration (using utterances of French sentences by an English speaker). The evaluation procedure involves aligning the automatic segmentation of each test utterance with a hand-segmentation, and computing an "alignment distance", which takes account of substitutions, deletions and insertions as well as the offsets in time between corresponding manual and automatic segment boundaries.

The procedure developed for evaluating this segmenter should be applicable to the evaluation of speech segmenters generally.

Section 2 of this paper gives a description of the SPELL segmenter. Section 3 describes the evaluation procedure. Section 4 gives details of the experiments and reports the results obtained; and section 5 contains a summary of conclusions from the study.

2. THE SPELL SEGMENTER

The application for which the SPELL segmenter is intended is the location (and, where variant pronunciations are possible, identification) of phonetic segments in an utterance of a known phrase or sentence by a non-native speaker of a language. Once a segmentation has been obtained, it is used, together with the results of other analysis, to assess the speaker's pronunciation.

The accuracy of segmentation required will depend on what aspect of pronunciation is being measured. For instance, for intonation assessment, a pitch contour is extracted from the utterance and is aligned with the segmentation. In this case, fine details of segment boundary location are largely irrelevant: it is only necessary that each syllable be located in the correct region so that the correspondence of pitch movements to syllables is correctly determined. To assess other aspects of pronunciation, however, such as vowel reduction, a more accurate segmentation may be required.

The segmenter has to run in near real time on a personal computer. Therefore computational efficiency was a major consideration in its design. Accordingly a simplified form of hidden Markov

model representation was adopted, in which each of a suitably defined set of acoustic-phonetic units (APUs) is represented by a one-state model with a single cepstral centroid vector.

The APUs are a mixture of phones and portions of phones, together with one unit corresponding to silence. The rationale of the APU set design is that each APU should be representable, to a first approximation, by a steady state with a single cepstral target. To achieve this, phones with definite temporal structure — stops, affricates and diphthongs — are split into smaller units.

The APU models are trained on segmented speech. The acoustic representation used consists of the 0th to 9th linear predictive cepstral coefficients [2] computed in a 20ms window every 5ms, with the 0th coefficient multiplied by an empirically determined scaling constant (here set to 0.25). For each APU, the cepstral centroid is obtained by averaging together the weighted-mean vectors for all training segments of the appropriate APU identity — where the weighting within each segment is by a raised cosine function so that the vectors near the centre of the segment are given most weight. The other parameters estimated for each model are a scale factor for the cepstral distances (which is made inversely proportional to an estimate of the mean squared Euclidean distance from the centroid); a self-transition probability; and a gamma distribution for segment duration.

The possible pronunciations of the phrase to be segmented are represented by paths through a *phrase model* or pronunciation network. This is derived automatically from an orthographic transcription of the phrase together with phonemic representations of the words (which may include multiple pronunciations). Phonological effects such as optional assimilation and reduction at word boundaries can be incorporated. The network is minimised in the sense that nodes with the same predecessor or successor set and the same APU label are conflated.

The segmentation proceeds using a Viterbi algorithm [3], in which the best-matching APU sequence from the phrase model is found together with its best-scoring alignment to the cepstral vector sequence representing the utterance. For each APU, the scaled squared Euclidean distance between the centroid and the observed vector for the current frame of the utterance is taken as the negative log emission probability (this corresponds to using a simple form of multivariate Gaussian distribution), and the negative log self-transition probability is taken directly from the model. When a transition between APUs is made, an adjustment is made to the accumulated negative log probability to convert the exponential duration probability distribution for the current APU, implicit in the basic HMM formulation, to the estimated gamma distribution. Such a postprocessing form of duration modelling has been found to give similar results to an exact formulation, with much less computation [3]. The implementation adopted imposes a maximum duration on each APU. Silence, however, is represented by an unconstrained-duration model, in which the segment duration is unlimited and the transition and duration log-probability terms are set to 0.

The segmenter's output is a series of segments, each with start and end times and an APU label.

3. EVALUATION PROCEDURE

The quality of the segmentations produced by the automatic segmenter is evaluated by comparing them with hand-segmentations of the same utterances.

The comparison of manual and automatic segmentations is complicated by the fact that there are different possible realisations of a given phrase as an APU sequence, and so the two segmentations

was conducted in two configurations: configuration A, with em1 as the seed speaker and em2 as the test speaker, and configuration B, in which the speakers' roles were interchanged. Within each configuration, the procedure was as follows.

Firstly, the seed speaker's enrolment utterances were hand-segmented, and seed APU models were trained on the segmented data. Next, these seed models were used, together with appropriate general (multiple-pronunciation) phrase models, to segment the test speaker's enrolment utterances. APU models for the test speaker were trained using the segmented enrolment utterances. (The details of this process were varied from one experiment to another, as described in section 4.2 below.) Then the test speaker's APU models were used, again in conjunction with general phrase models, to segment this speaker's utterances of the test sentences. These utterances were also segmented by hand, and the hand-segmentations and automatic segmentations were compared.

In the inter-language evaluation, em2 acted as the test speaker, with em1 and the French male speaker fm1 as seed speakers. Models for the English APUs were derived as before, using em2's enrolment utterances and em1's seed models. The models for the French APUs were derived by a similar procedure, using a set of 22 French enrolment sentences (designed to contain at least 5 occurrences of each of the 43 French APUs), with seed models derived from hand-segmented utterances of these same sentences by fm1. The evaluation was done on a set of 100 French sentences spoken by em2. The phrase models used in the segmentation of em2's French utterances included both French and English APUs, to allow for possible mispronunciations — hence the need for English APU models as well as French ones.

4.2 Results of intra-language experiments

A series of experiments was performed to optimise the segmentation. The results of some of these experiments are shown in table 1. (The other experiments involved adjustments to various numerical parameters in the APU model training, none of which resulted in any major improvement to the segmentation performance.) Each experiment is characterised in terms of the APU models used in segmenting the test utterances. The numbers given under "A" and "B" are the average alignment distances per test utterance in the two evaluation configurations.

The first comparison made was between using the seed models trained on the other speaker's data (experiment 001) and using the models for the current speaker trained on the seed models' segmentations of the enrolment utterances (experiment 002). In configuration B, the models trained for the current speaker gave the better segmentations. In configuration A, however, the segmentation performance got worse after enrolment. This can be attributed to the inaccuracy of the segmentations of the enrolment utterances based on the seed models: although the cepstral centroids in the "002" models are better for the current speaker than those in the seed models, their other parameters are badly affected by the poor initial segmentations.

In experiment 015, models were trained as in experiment 002, but the duration distributions (and self-transition probabilities) were then replaced by those from the seed models, which were more reliable, being derived from hand-segmentations. This improved the performance substantially. In experiment 018, the cepstral-distance scale factors were also taken from the seed models; this further improved the performance on the test data.

In experiment 016, the models from experiment 015 were used to resegment the enrolment utterances, and new models were trained on the segmentations; again the duration distributions from

to be compared may differ not only in their placement of segment boundaries but also in the number of segments and the sequence of APU labels attached to them. (To avoid this, it would be necessary either to force the hand-segmentation to use the same APU sequence chosen by the automatic segmenter (inconvenient for the human segmenter, and too lenient when the APU sequence chosen is inaccurate) or else to force the automatic segmenter to use the APU sequence found in the hand-segmentation (unrealistic in that a hand-picked APU sequence would not be available in the real application, where the segmenter has to select a pronunciation from a general phrase model).) A comparison procedure is therefore needed which can cope with APU substitutions, deletions and insertions, as well as with offsets in time between corresponding boundary placements.

The comparison problem has been solved by the implementation of a dynamic programming alignment procedure incorporating penalties for boundary offsets, substitutions, deletions and insertions. The penalty for a boundary matching is proportional to the squared offset between the boundary times in the two segmentations. Penalties can be specified for substitutions, deletions and insertions of particular APUs, or classes of APUs, as well as default penalties for substitutions, deletions and insertions not listed explicitly.

The alignment of a pair of segmentations consists of a sequence of segment boundary matches interspersed with APU substitutions (including matchings of identical APUs), deletions and insertions. The first boundary match aligns the start times of the two segmentations, and the last aligns their end times. The structure of a segmentation alignment can be specified formally as

$$B\ (\ [\ \{\ I^*\ |\ D^*\ \}\ S\ ]\ \{\ I^*\ |\ D^*\ \}\ B\ )^*$$

— where B represents a boundary match, D a deletion, I an insertion and S an identity or non-identity substitution; an asterisk denotes 0 or more occurrences; anything in square brackets is optional; and "$\{\ X\ |\ Y\ \}$" means "X or Y". The algorithm finds the alignment with the smallest total penalty; this total penalty is referred to as the *alignment distance* for the pair of segmentations being compared. Detailed evaluation statistics can be extracted from the alignments.

This evaluation procedure represents an advance over the segmenter evaluation previously used at CSTR [4] in that it does not require the APU sequences in the manual and automatic segmentations to be identical.

4. EXPERIMENTS AND RESULTS

4.1 Experimental procedure

It is envisaged that in practice the segmenter will be applied to utterances (of teaching or test sentences) by the current speaker, with APU models trained on this speaker's prior enrolment utterances. During the speaker-specific model training, the enrolment utterances will be segmented using a set of seed models derived from hand-segmented data from some other speaker or speakers. The evaluation experiments were designed accordingly.

For English seed model training and enrolment, a set of 22 sentences was defined. These sentences were designed to contain, in their probable pronunciations, at least 5 occurrences of each of the 61 English APUs. A disjoint set of 100 sentences was used in testing the segmenter.

For the intra-language evaluation, utterances of both enrolment and test sentences were recorded from two male RP English speakers (em1 and em2). To make full use of the data, the evaluation

Table 1: Results of intra-language experiments

Experiment: code	APU models	alignment distance A	B	average
001	seed models	3.356	2.902	3.129
002	trained on initial segmentations of enrolment data	3.892	2.192	3.006
015	centroids and scale factors as 002, durations from seed models	2.369	1.859	2.114
018	centroids as 002, scale factors and durations from seed models	2.238	1.651	1.944
016	as 015 with one iteration on enrolment data	2.234	1.758	1.996
019	as 018 with one iteration on enrolment data	1.977	1.544	1.760
025	as 019, scale factors and durations reestimated after iteration	2.412	1.455	1.934
026	as 019, scale factors reestimated after iteration	2.029	1.433	1.731
020	as 018 with two iterations on enrolment data	2.002	1.555	1.778
022	as 020, scale factors and durations reestimated after iterations	2.276	1.507	1.882
023	as 020, scale factors reestimated after iterations	2.285	1.375	1.830
027	trained on hand-segmented enrolment data	1.891	1.439	1.665

the seed models were substituted in. A similar procedure was followed in experiment 019, where the models from experiment 018 were used, and the scale factors and duration distributions were taken from the seed models. In each case the iteration improved the models. When the cepstral distance scale factors and (optionally) duration distributions were taken from the enrolment data after the iteration, the models were improved slightly in configuration B, but worsened in configuration A (experiments 025 and 026).

A second iteration (experiment 020) and reestimation of the duration distributions and scale factors following it (experiments 022 and 023) resulted in no further improvement overall.

In experiment 027, the models were trained using the hand-segmentations of the test speaker's enrolment utterances. The results were not much better than those of experiments 019 and 026 — indicating that only modest improvements could be expected from any refinement of the procedure for automatic segmentation of the enrolment utterances.

The main conclusions from these comparative experiments are that only the cepstral centroids, and not the cepstral-distance scale factors and duration distributions, should be reestimated after segmentation of a new speaker's enrolment utterances with the seed models; and that the enrolment utterances should be resegmented with the resulting speaker-specific models, and the cepstral centroids reestimated, to obtain an improved set of APU models for the new speaker.

The segmentations obtained in experiment 019 were examined in more detail, to find any particular problems which might be alleviated in future development of the system.

First, the alignment program was run again, with the penalties for boundary offsets made smaller to ensure that corresponding parts of the APU sequences were matched together even where

Table 2: Summary statistics for segmentations in intra-language experiment 019

statistic	A	B	overall
mean signed offset in frames (and in milliseconds)	1.2 (6.1)	-0.5 (-2.7)	0.3 (1.7)
mean magnitude of offset in frames (milliseconds)	3.0 (15.0)	2.7 (13.5)	2.8 (14.2)
APU substitutions per utterance	1.64	1.29	1.465
deletions per utterance	0.93	0.60	0.765
insertions per utterance	0.81	0.64	0.725

they were badly misaligned in time. Information on boundary offsets, substitutions, deletions and insertions was derived from the alignments. Some overall statistics are given in table 2. All instances of particularly inaccurate boundary placement (with offsets of 100ms or more in either direction) were examined individually. It was found that 14 of the 49 large offsets occurred near the ends of utterances (after the vowel in the final syllable), where the task of segmentation was made more difficult by prepausal lengthening, reductions in amplitude and post-utterance noise; 3 of the others were localised errors in boundaries between voiced sounds; and the remaining 32 instances occurred in 13 separate sequences of boundaries, with lengths ranging from 1 to 9.

Possibly more important than the absolute magnitude of the offset in an inaccurate segmentation is whether it results in any APUs' being mapped to locations which do not overlap their true locations. (This may happen even for a moderate offset if the segments are short; and it may not happen even for a large offset if they are very long, as is common at the end of an utterance.) Accordingly, a further examination was made of all instances where the segment found by the segmenter did not overlap or abut the corresponding segment in the hand-segmentation. The most prominent feature of the results was that there was a strong tendency for stop releases to be placed later in time in the automatic segmentations than in the hand-segmentations. This could be because the cepstral-distance scale factor was usually lower in the closure model than in the release model, resulting in a preference for extending the closure segment to include the release.

Some segmentation errors appeared to be due to the absence of the actual pronunciations (as shown in the hand-segmentations of the test utterances) from the phrase models. In some cases this was because there were unanticipated assimilation effects. There were also occasionally utterance-internal pauses, which were not allowed by the phrase models.

4.3 Results of inter-language experiments

A series of experiments like those for the intra-language evaluation was performed in the inter-language configuration. This time the phrase models included optional silence segments between words. One additional experiment (028) was performed; but there was no counterpart to experiment 027 since the enrolment utterances had not been hand-segmented. The results are shown in table 3.

The pattern of the results is broadly similar to that for the intra-language experiments. The best result (a mean alignment distance of 3.508) was obtained with two iterations of the enrolment procedure and no reestimation of the scale factors and duration distributions. It seems from these results that a larger number of iterations on the enrolment data (two instead of one) may be optimal in the inter-language case than in the intra-language case; however, some caution is

Table 3: Results of inter-language experiments

Experiment: code	APU models	alignment distance
001	seed models	4.228
002	trained on initial segmentations of enrolment data	5.196
015	centroids and scale factors as 002, durations from seed models	4.241
018	centroids as 002, scale factors and durations from seed models	3.898
016	as 015 with one iteration on enrolment data	4.343
019	as 018 with one iteration on enrolment data	3.636
025	as 019, scale factors and durations reestimated after iteration	4.704
026	as 019, scale factors reestimated after iteration	3.887
020	as 018 with two iterations on enrolment data	3.508
022	as 020, scale factors and durations reestimated after iterations	4.570
023	as 020, scale factors reestimated after iterations	3.941
028	as 018 with three iterations on enrolment data	3.576

in order, given that the evaluations were on only two speakers in one case and one speaker in the other. The worsening of performance when the scale factors and duration distributions are reestimated is more marked in the inter-language results than in the intra-language case.

Statistics of the segment boundary offsets, APU substitutions, deletions and insertions for experiment 020 were computed. The mean signed boundary offset was -0.7 frame (-3.6ms), and the mean absolute offset was 2.6 frames (13.0ms). The rates of substitutions, deletions and insertions per utterance were 8.27, 0.84 and 1.50 respectively.

A comparison with the intra-language statistics in table 2 shows that, despite the higher average alignment distance in the inter-language case (3.508 per utterance, against 1.977 for the same test speaker in intra-language experiment 019), the boundary positioning is no less accurate on average and there are no more deletions. The higher distances result mainly from the large numbers of substitutions and (to a lesser extent) insertions. These phenomena were explored by examination of the particular substitutions and insertions occurring in the inter-language evaluation. It was found that, of the 827 substitutions, 409 were substitutions of a French APU for the corresponding English APU or vice versa — where "corresponding" means "represented by the same SAMPA symbol", which usually indicates close phonetic similarity. (In 376 instances the APU was a stop closure or release.) These substitutions do not, however, contribute much to the alignment distances, since their penalties are set to small values in the alignment parameter file. Of the remaining 418 substitutions, 304 were vowel substitutions. Often an insertion error resulted from the absence of the APU sequence in the hand-transcription from the phrase model.

A detailed examination was made of all segment boundary offsets of 100ms or more. There were 36 of these, of which 13 were early placements of utterance-initial boundaries before consonants (mostly stops) — perhaps resulting from a mismatch in the recording conditions for the French

and English enrolment utterances; 2 others (the 2 largest positive offsets) were in a final syllable; and the other 21 comprised 10 sequences of consecutive boundaries in 9 different utterances.

5. CONCLUSIONS

The main conclusions from this study are as follows.

Firstly, the SPELL segmenter appears to be accurate enough for the application primarily in view (intonation assessment) on a large majority of utterances. In the intra-language evaluation, apart from a few outliers (representing about 1% of the boundaries), the distribution of segment boundary offset magnitudes is very similar to that for a more sophisticated and computationally complex segmenter using hidden semi-Markov models trained on 200 hand-segmented utterances [4] — though the comparison is not exact since there were several differences between the two evaluations, including the use of different data. The most noticeable difference between the intra-language and inter-language results is in the numbers of APU substitutions and insertions; this is as might be expected, given the wider range of possible pronunciations (incorporating both native-language and foreign-language APUs) from which the segmenter has to choose in the inter-language case. Some very large boundary offsets occurred in the inter-language evaluation, but many of these were in initial boundaries between silence and a stop closure, where they are unlikely to do much harm. Ignoring initial-boundary inaccuracies, 90% of the French utterances were segmented with no offsets of 100ms or more from the hand-segmentation boundaries.

Secondly, to attain this level of performance, an enrolment procedure should be adopted in which the duration distributions and cepstral-distance scale factors from the seed models (estimated from hand-segmented data) are retained while the cepstral centroids are reestimated on the student's enrolment utterances, and the enrolment segmentation should be iterated once or twice to obtain improved cepstral centroids.

Thirdly, the procedure developed for evaluation of this segmenter, in which a dynamic programming algorithm is used to align each automatic segmentation with the corresponding hand-segmentation, is a useful tool which could be applied to the evaluation and optimisation of speech segmenters generally.

6. REFERENCES

[1] J-P Lefèvre *et al.*, "Macro and micro features for automated pronunciation improvement in the SPELL system", *Speech Communication*, vol.11, pp.31-44, 1992.
[2] J D Markel and A H Gray, *Linear Prediction of Speech*, Springer-Verlag, 1986.
[3] L R Rabiner, B-H Juang, S E Levinson and M M Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities", *AT&T Tech. J.*, vol.64, pp.1211-1234, 1985.
[4] M S Schmidt and G S Watson, "The evaluation and optimization of automatic speech segmentation", *Proc. Eurospeech 91*, pp.701-704.