Precision of Phoneme Boundaries Derived using Hidden Markov Models

Ladan Baghai-Ravary, Greg Kochanski, John Coleman

Phonetics Laboratory, Oxford University

ladan.baghai-ravary@phon.ox.ac.uk

Abstract

Some phoneme boundaries correspond to abrupt changes in the acoustic signal. Others are less clear-cut because the transition from one phoneme to the next is gradual.

This paper compares the phoneme boundaries identified by a large number of different alignment systems, using different signal representations and Hidden Markov Model structures. The variability of the different boundaries is analysed statistically, with the boundaries grouped in terms of the broad phonetic classes of the respective phonemes.

The mutual consistency between the boundaries from the various systems is analysed to identify which classes of phoneme boundary can be identified reliably by an automatic labelling system, and which are ill-defined and ambiguous.

The results presented here provide a starting point for future development of techniques for objective comparisons between systems without giving undue weight to variations in those phoneme boundaries which are inherently ambiguous. Such techniques should improve the efficiency with which new alignment and HMM training algorithms can be developed.

Index Terms: phoneme alignment, labelling accuracy, mutual consistency, data-driven analysis, hidden Markov models.

1. Introduction

Almost all current acoustic-phonetic speech technology (speech recognition, understanding, alignment and synthesis systems) is based on the concept of speech as a sequence of discrete speech units (phonemes). However in practice most realisations of these units are far from discrete – they blend into one another, so all boundaries are arbitrary, to a greater or lesser degree.

In the development of any new speech system, one of the first steps generally involves analysing a large database to identify the locations of individual phonemes or phoneme groups. The database being analysed is often prohibitively large for manual labelling, so an automatic system is required, but the design of such an automatic system is problematic because it is not clear how to assess its performance.

The most common method of assessing labelling accuracy is to compare the automatic system's labels with a 'gold-standard' set defined by one or more phoneticians [1, 2]. This is not only inherently subjective, but also very 'labour intensive' to produce. To avoid these difficulties, an objective measure of the quality of the alignment system is needed, without reference to human decisions. The measure developed and presented here quantifies inconsistencies between systems (where the time difference between the different systems' labels is unpredictable), but is independent of any *consistent* differences (where one system reliably places a particular label earlier or later than another).

In general, alignment precision varies as a function of the phones that define the boundary [1, 3]. Some such boundaries can be accurately predicted in one system if their position is known in another, while others vary greatly and are essentially unpredictable. This system-to-system predictability is the focus of this paper.

The approach described here bears some similarity to that of Kominek and Black [4], but it does not average the results of alignment systems to obtain a definitive set of labels – it compares every individual alignment system's result with every other one's and assesses how *precisely* the labels can be located – not *where* they should be. This allows for the fact that some alignment systems may tend to label a particular type of transition early or late relative to another. Such consistent discrepancies do not indicate a significant difference in precision between alignment systems, and so do not affect the results of this analysis.

2. Method

The approach here is to use a cohort of results from automatic systems to estimate the variability of individual labels. By building many different phonetic alignment systems with diverse characteristics and comparing the individual labels between these systems, it is possible to determine which boundaries are ambiguous and which can be defined accurately, without reference to human labels.

2.1. Phoneme Grouping

There are very many possible phoneme-to-phoneme boundaries (typically around 2000, depending on the details of the phoneme inventory). This is too many for simple interpretation of any analysis based on individual phoneme identities.

Furthermore, many of these phoneme-pairs are very rare in natural speech, and any statistics derived from such small numbers of examples would be unreliable. To avoid these problems, phonemes can be grouped into broad phoneme classes before calculating any statistics (see Table 1, which also shows the abbreviations for each class, as used in the rest of the paper).

Table 1. Broad phoneme classes (SAMPA [5]).

Class	Abbr.	British English Phonemes (SAMPA)
Nasal	Nas	m, n, N
Plosive	Plo	b, d, g, k, p, t
Affricate	Aff	dZ, tS
Fricative	Frc	D, S, T, Z, f, h, s, v, z
Vowel	Vow	@, A, E, I, O, U, V, {, 3, i, u, Q
Approximant	App	r, j, l, w
Diphthong	Dip	I@, U@, aI, aU, E@, eI, OI, @U
Silence	Sil	silence, short inter-word pause

2.2. Training Data

The training data for all systems was an *ad hoc* corpus originally assembled for other purposes. The subjects were all speakers of Standard British English, and the utterances were a mixture of single words, phrases and complete sentences of varying lengths. The recordings were made with different equipment and at different sampling rates, digitally re-sampled to 16 kHz. The database consists of over 23,000 utterances, making a total of 48,000 spoken words taken from a vocabulary of 16,000.

The phoneme strings to be aligned with the speech were based on a lexicon compiled from several sources, with manually edited transcriptions for any words in the database but not in any of the lexica. For sentences and phrases, an optional 'short-pause' label was inserted between words. No post-lexical rules were applied, so some of the phoneme transcriptions may have been phonetically unrealistic.

2.3. Alignment Systems

All the alignment systems were based on Gaussian mixture Continuous-Density Hidden Markov Models (CD-HMMs). The models and most of the pre-processing was performed with the HTK Hidden Markov Model Toolkit [6]. All the systems were based on the same lexicon and were trained on the same data, and with the same procedures and training parameters. The details which were changed between systems are as follows:

2.3.1. Signal Analysis

To ensure a reasonable amount of diversity in the alignments produced by the different systems, three different preprocessors were used to produce observation vectors:

1. Mel-Frequency Cepstral Coefficients (MFCCs) [7] with a 20 ms time window for each frame, and one frame every 10 ms.
2. Linear Prediction Cepstra (LP-Cepstra) [8] with a 24 ms time window, one frame every 8 ms.
3. Auditory Description Vectors (ADVs) [9], with a frame rate of one every 10 ms.

2.3.2. HMM Structure

The numbers of states, and of Gaussian mixtures per state, were the same for all phoneme models within each system, but were varied between systems to produce different alignment results.

The number of states per phoneme was varied between 2 and 5. For 2 and 3 state models, a strict left-right topology was used, with no skips. For 4 and 5 state models, 1-state skips were allowed so that shorter phonemes could still be modelled without extending the states of one model into neighbouring phonemes. The number of Gaussian mixtures per state was set to either 1, 2, 4, or 8 for all phoneme models in each system.

Thus for each of the 3 pre-processors, there are 4 different numbers of states per model, and 4 different numbers of mixtures per state, making a total of 3 × 4 × 4 = 48 alignment systems.

2.3.3. Training Procedure

The training process used embedded re-estimation via the Baum-Welch algorithm [10], applied in three phases:

- Training from flat-start HMMs, initialised to the global means of all the training data, to produce single-mixture phoneme, silence, and short-pause models.
- Disambiguation of alternative pronunciations (including presence or absence of inter-word pauses) followed by re-training of the models.
- Disambiguation as before, and an increase in the number of mixtures in appropriate states (using a randomised duplication of each existing mixture), followed by final re-training of the full models.

Four Baum-Welch iterations were performed at each stage.

2.4. Comparisons

The phonetic labels for each system were generated by conventional Viterbi alignment of the phonetic HMMs with the respective speech observation vectors. Each label from each alignment system was then compared with the equivalent label from each of the other systems. For each system-pair, the time offset between the labels was calculated. For the broad-class phoneme experiments, these were grouped according to the broad class of the phonemes involved. For brevity these groups will be referred to as *class-transitions* because they correspond to the transition from one broad phonetic class to another. The transitions in the other, ungrouped, experiments will be referred to as *phoneme-transitions*.

For each combination of class / phoneme-transition and system-pair, histograms were constructed to show how frequently each time discrepancy was observed. The histogram bins were set to 10 ms width (the worst-case time resolution obtainable with the front ends used here). Some typical histograms are shown in Figures 1 to 3.

Figure 1 shows the class-transition time-difference distribution which is asymmetrical, but where the system-pair is actually in good agreement (one labels the transitions 32 ms earlier than the other, but that difference is nearly constant, so the systems are mutually consistent). Figure 2 shows the distribution for a transition which is labelled inconsistently by the two systems, so the histogram is too broad for the systems to be considered a “match”. Finally, Figure 3 is a multi-modal distribution which, again, does not indicate a realistic agreement between the systems.

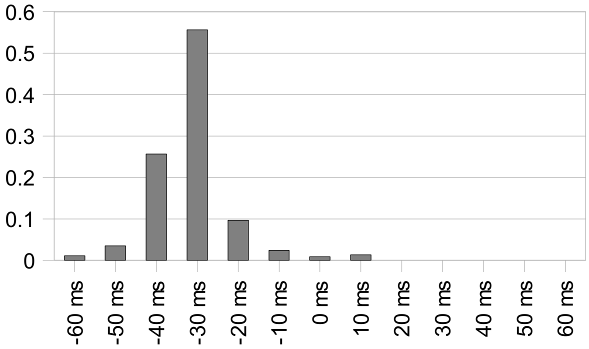

Figure 1: Normalised time discrepancy histogram for Sil → Nas transitions, comparing a system using ADV feature vectors, 2 states per model, 1 mixture per state, *vs*. LP Cepstra, 3 states, 4 mixtures

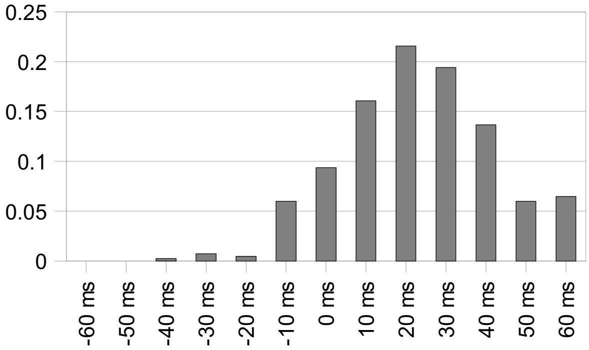

Figure 2: Normalised time discrepancy histogram for App → Sil transitions, comparing ADVs, 2 states, 1 mixture *vs*. LP Cepstra, 3 states, 8 mixtures

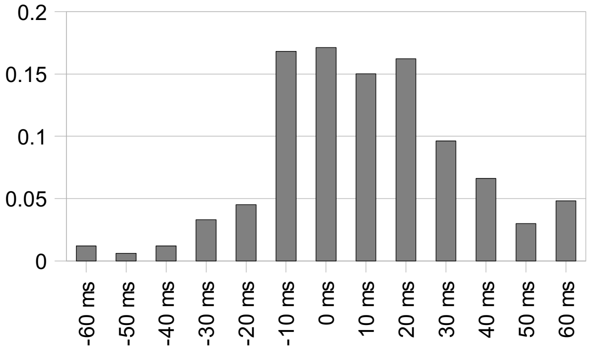

Figure 3: Normalised time discrepancy histogram for Dip → Sil transitions, comparing ADVs, 2 states, 1 mixture, *vs*. LP Cepstra, 2 states, 1 mixture

In order to decide which class-transitions are reliable and which are not, the respective histogram is examined, and if more than 75% of the discrepancies fall in two adjacent bins, the two alignment systems are deemed to agree. This is quite a strict criterion – they must agree to within their characteristic time resolution on the vast majority of each transition.

Two bins are used in the calculation since the peak of the underlying continuous distribution might lie close to the edge of a bin, in which case the magnitude of the true peak would

Table 2 Number of agreeing system-pairs for each class-transition

Transitions Class A → Class B		Class B							
		Plo	**Aff**	**Frc**	**Nas**	**App**	**Vow**	**Dip**	**Sil**
Class A	**Plo**	19	39	70	84	166	411	469	57
	Aff	369	–	25	–	63	688	753	123
	Frc	438	86	22	456	133	423	537	116
	Nas	241	632	247	5	52	332	396	84
	App	208	156	251	60	5	26	9	103
	Vow	332	365	254	182	8	1	1	78
	Dip	361	437	260	170	6	0	4	77
	Sil	923	395	299	817	713	513	516	–

be divided between the two bins, and so would not be detected at its true magnitude.

This histogram-based approach is used because it makes no strong assumptions about the distribution of values (which does not need to be Gaussian, uni-modal, or even continuous). Even if one system of each pair consistently places its labels earlier or later than the other, they are not penalised. As long as the discrepancies are consistent, they will still be identified as agreeing with each other.

Finally, the overall consistency with which any given class-transition is identified is quantified by counting how many system-pairs agree. In these experiments there are 48 different systems so there are (48 × 47) / 2 = 1128 possible pairs. The more pairs agree, the more reliable the labelling of that class-transition.

3. Results

The broad-class results are shown in Table 2, with the background shading denoting the magnitude of the respective values (to make the structure in the figures more clear). The darker the cell, the more reliable the class-transition. There are some values missing from Table 2. There were no occurrences of an affricate-to-affricate transition in the data used for these experiments, and there were insufficient examples of affricate-to-nasal transitions to calculate meaningful statistics. A silence-to-silence transition would be nonsensical.

The corresponding table for the ungrouped phoneme-transitions would be too large to present here, but they have been compared to the grouped results and some of the more interesting observations are summarised below.

3.1.1. Utterance-Initial Transitions

On average, 72% of the system-pairs agree on silence-to-phoneme transitions, but a significant number of silence-to-phoneme transitions were identified identically by 99% or more of all system-pairs:

silence → /S/, /d/, /b/, /w/, /l/, /r/, /m/, /n/, or /O/

Conversely, the least reliable silence-to-phoneme transitions (agreed upon by 40% or less of system-pairs) are:

silence → /v/, /T/, /h/, /f/, or /tS/

3.1.2. Utterance-Final Transitions

On average, only 13% of the system-pairs agree on phoneme-to-silence transitions, and only the following transitions are consistently identified by more than 25% of system-pairs:

/s/ or /S/ → silence

The following utterance-final transitions are the ones identified least accurately, with less than 5% of system pairs agreeing on their timings:

/k/ or /v/ → silence

3.1.3. Most Reliable Transitions

A number of phoneme-transitions are agreed upon by over 99% of the system-pairs. They are:

/s/ → /@U/, /eI/, /aI/, /O/, /E/, /u/, /i/, /V/, or /p/
/t/ → /eI/, /u/, or /i/
/f/ → /A/
/z/ → /eI/

The vast majority of these correspond to a change in excitation from either a fricative or voiceless stop to a voiced phoneme (vowel or diphthong).

3.1.4. Most Unreliable Transitions

The following phoneme-transitions are only consistent between 1% or less of the system-pairs:

/U@/, /O/, /E/, /@/, /u/, or /i/→ /l/
/aI/ or /aU/ → /@/
/j/ → /u/
/U@/ → /r/
/l/ → /eI/, or /w/
/s/ → /s/
/t/ → /t/

Most of these involve transitions between vowels, diphthongs and approximants.

It should be borne in mind that the /s/ → /s/ and /t/ → /t/ transitions are especially problematic because they are not generally articulated as two distinct phonemes. Most frequently they are assimilated into a single sound. This is a problem with the simple lexicon-based approach used both in this paper and in most current speech recognition systems.

4. Discussion

An important feature of the analysis is that it allows systematic differences to be separated from unpredictable alignment errors. For instance, Figure 1 shows a pair of alignment systems that have a large systematic difference in where they put a phoneme boundary, but they disagree by a consistent amount. Such systematic disagreements can be as large as (or sometimes larger than) the unpredictable alignment errors. These systematic errors reflect the fact that phoneme boundaries are ill-defined objects and that each alignment system identifies them in its own characteristic way. Systematic differences are not actually problematic or "wrong".

Table 2 suggests that the most reliable transitions are the ones from silence to the other phonetic classes. In particular, the most reliable of all are silence-to-plosive and silence-to-nasal. Conversely, the least reliable are between vowels, approximants and diphthongs. The transitions to silence are also relatively unreliable and variable. These conclusions are supported by the results obtained for individual phonemes.

The range of agreement scores is remarkably large: some phoneme-transitions (e.g. silence → /n/) are consistent between 99% of system-pairs, while some (e.g. /O/ → /l/) are almost never consistent.

Table 3 summarises the ranges of reliability for different class-transitions. As would be expected, many of the most precisely identifiable class-transitions correspond to transitions from one form of excitation to another – voiced, fricative, affricate, stop, etc. – and most of the rest consist of abrupt changes to the vocal tract configuration – between nasals, vowels, and approximants, for example.

Table 3. Subjective transition groupings

Agreeing Pairs	Class-Transitions
299 – 923	Silence → All
600 – 800	Aff → Dip, Vow Nas → Aff
400 – 600	Frc → Plo, Nas, Dip, Vow Plo → Dip, Vow Dip → Aff
300 – 400	Aff, Vow, Dip → Plo Nas → Vow, Dip Vow → Aff
200 – 300	App, Nas → Plo App, Nas, Dip, Vow→ Frc
100 – 200	App → Aff Frc, Plo → App Dip, Vow → Nas
0 – 100	App, Dip, Vow → App, Dip, Vow App, Nas → App, Nas Plo → Plo, Nas Plo, Aff, Frc → Aff, Frc Aff → App
53 – 123	All → Silence

These discrepancies observed between system-pairs are broadly comparable with those between human segmenters, as documented in the literature [1, 3]. The pattern of these human errors is broadly similar to the those reported here, but with some notable discrepancies.

An inverse relationship would be expected between mean human segmentation error and the agreement scores presented here. This is supported for extreme pairs; for instance the classes with the three largest human errors (Nas→Nas, Vow→Vow, and Vow→App) have very low agreement scores in our work, and two classes with the smallest human errors (Frc→Plo and Plo→Vow) are among our best-agreeing pairs. However, other pairs disagree.

Some of these differences may be attributed to the shortcomings of the lexicon-based approach used in this paper (which does not always yield the correct pronunciation of any given utterance), but this is an area which is the subject of further investigation.

5. Conclusions

While all phoneme boundaries are somewhat artificial and perhaps even arbitrary, some are more clearly ambiguous than others. Consequently different boundaries should be given different weights when assessing the accuracy of any alignment or segmentation.

The results presented here have been derived without reference to any human alignments, but they broadly confirm independently-assessed human labelling variations. They clearly indicate that the boundaries between vowels, diphthongs and approximants are highly ambiguous, and so do not provide an effective way of comparing alignment systems. Word endings (transitions from any phoneme to silence) are also relatively unreliable, and of little use in assessing accuracy.

On the other hand, word onsets (i.e. silence-to-phoneme transitions) are clearly identifiable almost regardless of the phoneme at the start of the word, and so are critical in assessing accuracy. Other reliable transitions include affricates-to-vowels / diphthongs, and nasals-to-affricates. The most precisely identifiable class-transition of all is the silence-to-plosive.

Ultimately these observations should not only affect the methods used to optimise alignment systems; they should also be taken into account when designing speech databases for the training of speech recognition and synthesis systems.

6. Acknowledgement

This project is supported by the Economic and Social Research Council (UK) via project RES-062-23-1172.

7. References

[1] Wesenick, M.-B., and Kipp, A., "Estimating the quality of phonetic transcriptions and segmentations of speech signals", Proceedings ICSLP-96, pages 129–132. IEEE, 1996.

[2] Vonwiller, J., Cleirigh, C., *et al.*, "Development and application of an accurate and flexible automatic aligner", International Journal of Speech Technology 1, pages 151-160, 1997.

[3] Cosi, P., Falavigna, D., Omologo, M., "A preliminary statistical evaluation of manual and automatic segmentation discrepancies", in Proc. Eurospeech 1991, pp. 693-696.

[4] Kominek, J., and Black, A., "A family-of-models approach to HMM-based segmentation for unit selection speech synthesis", in Proc. Interspeech 2004, Jeju Island, Korea, 2004.

[5] Wells, J. C., "SAMPA computer readable phonetic alphabet". In Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV, section B.

[6] Young, S. J., *et al.*, "The HTK book". Cambridge University Engineering Department, December 2006.

[7] Bridle, J. S., and Brown, M. D., "An experimental automatic word recognition system," JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, England. 1974.

[8] Atal, B. S., "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", in J. Acoust. Soc. America, vol. 55, no. 6, pp.1304–1312, June, 1974.

[9] Kochanski, G., and Orphanidou, C., "Testing the ecological validity of repetitive speech". Proc. International Congress of Phonetic Sciences (ICPhS 2007), Saarbrücken, Germany. 10 Aug 2007.

[10] Baum, L. E., Petrie, T., Soules, G., and Weiss, N., "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", Ann. Math. Statist., vol. 41, no. 1, pp. 164-171, 1970.