Objective Optimisation of Automatic Speech-to-Phoneme Alignment Systems

Ladan Baghai-Ravary, Greg Kochanski, and John Coleman

Oxford University Phonetics Laboratory,
41 Wellington Square, Oxford, OX1 2JD, UK
ladan.baghai-ravary@phon.ox.ac.uk

Full reference: "Objective Optimisation of Automatic Speech-to-Phoneme Alignment Systems", L. Baghai-Ravary, G. Kochanski and J. Coleman, *Human Language Technologies as a Challenge for Computer Science and Linguistics;* Zygmunt Vetulani (ed.); pp. 341-5; 6-8 November, Poznan, Poland 2009; ISBN 978-83-7177-746-2. May be downloaded from http://kochanski.org/gpk/papers/2009/ltc-final-website.pdf .

Abstract

This paper presents techniques for objective characterisation of Automatic Speech-to-Phoneme Alignment (ASPA) systems, without the need for human-generated labels to act as a benchmark. As well as being immune to the effects of human variability, these techniques yield diagnostic information which can be helpful in the development of new alignment systems, ensuring that the resulting labels are as consistent as possible. To illustrate this, a total of 48 ASPA systems are used, including three front-end processors. For each processor, the number of states in each phoneme model, and of Gaussian distributions in each state mixture, are adjusted to generate a broad variety of systems. The results are compared using a statistical measure and a model-based Bayesian Monte-Carlo approach. The most consistent alignment system is identified, and is (as expected) in close agreement with typical "baseline" systems used in ASR research.

Keywords: phonetic alignment, label accuracy, phoneme detectivity, ASPA systems

1. Introduction

Manual labelling of large databases necessitates the use of teams of labellers, and the individuals within the team invariably differ in the detailed interpretation of any labelling guidelines they have been given (Lander, 1997). The effect of these differences could be minimised by getting a number of labellers to label the data independently, and then to take the median position (for example) of each label as being the "correct" value. This is impractical for databases of any significant size, and so it is preferable to automate as much as possible of the labelling process. Not only does this reduce the cost of the process, but it also has the potential to reduce inconsistencies in the labels, due to factors such as the subjectivity of visual perception of spectrogram data.

Automatic Speech-to-Phoneme Alignment (ASPA) systems are designed to respond consistently to well-defined features of the acoustic signal, and will always produce the same results, when presented with the same data. They are not influenced by the inconsistencies of human perception.

HMMs are often used for ASPA, and although each system is repeatable, each may focus on different aspects of the speech, resulting in discrepancies between the labels from different systems.

To identify which of these systems gives the most authoritative labels, we have compared a very large number of alignment results from different ASPA systems, and used these comparisons to evaluate each system's performance. Those labels which are positioned consistently with respect to those of other systems are deemed reliable and are then used to decide on the relative merits of the different systems as a whole.

2. Experiments

A total of 48 ASPA systems were built using the HTK HMM Toolkit (Young et al., 2006). The systems varied in the number of states per phoneme, the number of Gaussian components per mixture, and the initial front-end processing used to produce the observation vectors.

2.1. Data

The experiment used an ad hoc corpus assembled for other purposes. The 34 subjects were all speakers of Standard British English, and the utterances consisted of both single words and complete sentences. The recordings were made with different equipment and at different sampling rates, but all were digitally re-sampled to 16 kHz using the "rate" operation of SoX (SoX, 2009). Overall, the database consists of just over 23,000 utterances, making a total of 48,000 spoken words taken from a vocabulary of 16,000.

2.2. Front-end Processing

The three different pre-processors chosen were Mel-Frequency Cepstral Coefficients (MFCCs), Linear Prediction (LP) Cepstrum, and Auditory Description Vectors. All were computed at 12.5 ms intervals.

MFCCs are the de facto standard baseline for most speech recognition experiments. The HTK MFCC_D_0 implementation was used, i.e. 13 MFCC coefficients (including the zeroth energy coefficient) with 13 dynamic ("delta") coefficients, with an analysis window of 25 ms.

Linear prediction cepstral coefficients are broadly similar to MFCCs, but with frequency domain smoothing being an inherent part of the spectral analysis, rather than the perceptually based "filter-bank" smoothing of the

MFCC. We used the LPCEPSTRA_E_D front-end of HTK, giving 13 static coefficients, including the energy, together with 13 dynamic coefficients, again with an analysis window of 25 ms.

Auditory Description Vectors (ADVs), proposed in (Kochanski and Orphanidou, 2007), were designed to mimic human perception of phonemes. The vector used here is derived from Equation 2 of Kochanski and Orphanidou by a dimensionality-reducing linear transformation which yields 19 coefficients.

The transformation is designed to maintain the distance between phonetically dissimilar data in a fashion similar to Linear Discriminant Analysis, as described in (Sebestyen, 1962) and applied to auditory data in (Beet and Gransden, 1992).

The definition for the initial acoustic description vectors differs slightly from that in (Kochanski and Orphanidou, 2007): primarily, a shorter time domain smoothing window is used here. The vector consists of:

- The loudness in 1 erb-wide (Moore and Glasberg, 1983) bins, smoothed with a 60 ms wide $1+\cos^2$ window.
- Five broader-band spectral features computed from 20 ms windows.
- Five broad-band edge detectors, taking a difference across a 40 ms interval.
- Two entropy features. One is the entropy of the spectrum (computed over a 30 ms window). The other is a space-time entropy in a 70 ms window, which will also pick up changes from one frame to the next. The latter tends to be higher at phoneme edges and in rough voicing.
- Two voicing features. One derived from the entire signal, the other from the high-frequency parts (above 1 kHz).
- One "roughness" feature inspired by (Hutchinson and Knopoff, 1978). This is designed to reflect fluctuations in the neural firing rate on 3-30 ms time scales. It would be large if fed a pair of pure tones 50 Hz apart, but small if the tones were 1 Hz apart or 1000 Hz apart.
- Finally, one pseudo-duration feature (Kochanski et al., 2005) which has a variable window width, but one that is comparable to the phoneme duration.

This 51-dimensional representation is then reduced to 19 dimensions by retaining the largest terms in an eigenvalue expansion.

2.3. HMM Training

All the phoneme strings to be aligned with the speech were based on a lexicon compiled from diverse sources. An optional short-pause phoneme was added between words. No post-lexical rules were applied, so the phonetic transcriptions will have some inaccuracies.

All the aligners compared in this work were based on broadly similar Continuous-Density Hidden Markov Models (CD-HMMs), implemented using HTK. The number of states and the numbers of Gaussian mixtures in each state were varied to evaluate the effects of changing these parameters.

The training process was similar for all the experiments: it used "embedded re-estimation" using the Baum-Welch algorithm (Young et al., 2006), applied in three phases (Baghai-Ravary, 1995), with four training iterations at each stage.:

- 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 each state (using a randomised duplication of each existing mixture), followed by final re-training of the full models.

3. Differences Between Systems

The methods described here assess the relative performance of each alignment system in terms of how consistently the labels are positioned, relative to other occurrences of the same label identified by other systems. The word "label" here denotes the transition from one phoneme to another. The former will produce close to 2000 potential label identities, making it difficult to present and generalise the results.

By using broad phonetic classes instead of phonemes, the number of statistics can be reduced to 55, and the statistics themselves become more reliable, being derived from many more observations.

3.1. Pairwise Variance

The simplest approach to measuring differences between ASPA systems is to compute the mean and variance of corresponding labels for each pair of systems. If there are large mean differences between the labels produced by two systems, it does not necessarily imply that either system is poor, just that they have systematically different definitions for where one phoneme stops and the next one starts. It is only if the variance of the time differences is large that one system can be said to be inconsistent with the other. Thus the reliability of a set of HMM alignments can be assessed by the variances of the discrepancies in its labels with respect to the other systems.

Suppose that one had a "bad" alignment system where label position is very sensitive to small changes in the local acoustic properties. Contrast this with a "good" system where small changes in the acoustical properties lead to small changes in the label positions. When we calculate the differences between corresponding phonemes, we are essentially giving a random sample of acoustical properties to the systems. The variance of the difference will then be related to the strength of the relationship between acoustical properties and timing. Generally, greater sensitivity of either system will lead to a larger variance of the differences.[1] A high sensitivity of either system implies a large variance. Conversely, if a given system always yields

1 Potentially the two systems might have large but nearly identical dependences on the local acoustical properties, but it is unlikely, given that representations are many-dimensional, and therefore there are many ways to differ. Therefore this possibility is ignored.

a small variance when compared with other systems, it is likely to be "good" in the sense of having a relatively small sensitivity of phoneme boundary position to acoustic properties.

The variance observed for different phoneme-class to phoneme-class transitions can be quite different. Labels for vowel-to-liquid transitions, for example, often differ widely without being fundamentally "wrong", because human labellers treat such transitions as broad and unclear. To avoid the necessity of modelling these variances in detail, we order the different systems' results for each phoneme-class to phoneme-class transition according to their variances. Then, to decide which system is most consistent, we calculate the mean ranking of the system's phoneme-class to phoneme-class discrepancies against all other systems. In this way, we make no assumptions regarding the range of variances, or even the inherent linearity of the scale.

A low ranking implies that the system is generally consistent with many of the other systems and thus that the positions of the labels it generates are relatively insensitive to small changes in acoustic properties. Such a system presumably responds to the most reliably identifiable features of the acoustic signal. Conversely a system with a lot of random variability would have a high rank, suggesting that it responds to incidental acoustic properties.

3.2. Parametric Bayesian Models

Another, parallel analysis builds a model that predicts where a given aligner will mark a given boundary. We search for the parameters of this model that best explain the observed boundary positions; some of the parameters correspond to systematic differences from one system to the next; other parameters correspond to the consistency of a system.

3.2.1. Phoneme "Detectivity"

We assume that (for a given alignment system) each class of phonemes tends to be systematically longer or shorter than the overall average. This can be thought of as a class-dependent and aligner-dependent "detectivity". A given aligner tends to be more sensitive to certain phonemes: phonemes that an aligner detects more easily tend tend to expand in both directions; phonemes that are not easily detected by an aligner tend to have shorter lengths than nominal.

Qualitatively, imagine a slice of sound in an ambiguous region[2] between two phonemes: the sound is intermediate between the two, but whichever phoneme is easier for a given system to detect will win. Quantitatively, if C is the broad phonetic class of a phoneme, we write:

$$detectivity = d(C, aligner)$$

In this model, the boundary shift is:

$$\Delta t = \sigma_{ab} * (d(C_a, aligner) - d(C_b, aligner))$$

where σ_{ab} is the standard deviation observed between the labels for transitions from class C_a to class C_b. The result of this model is that phonemes with a larger detectivity for a given system will tend to be longer. Ambiguous boundaries will tend to have larger values of σ_{ab} and would be expected to have larger systematic shifts, Δt.

Another motivation for this model is that it expresses diphone properties in terms of phoneme properties. That means you only need c. 45 numbers to predict the boundary positions for all (over 1000) diphones. To the extent that it works well, it makes the overall results much easier to compute and understand.

3.2.2. Interpretation

This model predicts systematic differences between one aligner and another, but there are also disagreements between aligners that are not predictable from a knowledge of the two neighbouring phones. The standard deviation of these unpredictable disagreements is the σ_{ab} value mentioned above.

We assume that σ_{ab} can be approximated as:

$$\sigma_{ab} = K_\sigma(aligner) * K_C(C_a, C_b) * \sigma_{global}$$

so that some aligners are proportionally better than others (expressed by K_σ), and that some class-to-class boundaries are easier to mark than others (K_C).

The smaller K_σ is, the better the aligner. The bigger K_C is, the more difficult it is to define the boundary between two phoneme classes. The factor allows us to compare diphones. Finally, σ_{global} (which is a single, overall parameter) is the typical size of disagreements between aligners.

Overall, between detectivities and K_σ values, K_C values and σ_{global}, there are 144 parameters in our current analysis. To estimate these parameters we have used an adaptive Markov-Chain Monte-Carlo process (Kochanski and Rosner, 2009). This algorithm can provide statistically valid confidence intervals for its parameters.

4. Results

4.1. Pairwise Variance Results

For each front end, the optimum (i.e. most mutually-consistent) set of HMM training parameters was found using the approach described in section 3.1 above. The results for each front-end are shown separately below (Figures 1 to 3). In these figures, a rank of zero means that the respective ASPA system agrees more closely than any other with at least one system, for *all* broad-class phoneme transition labels.

2 The ambiguity can be either intrinsic in the sound or it could be due to a sharp acoustic boundary falling near the middle of a window used for front-end processing.

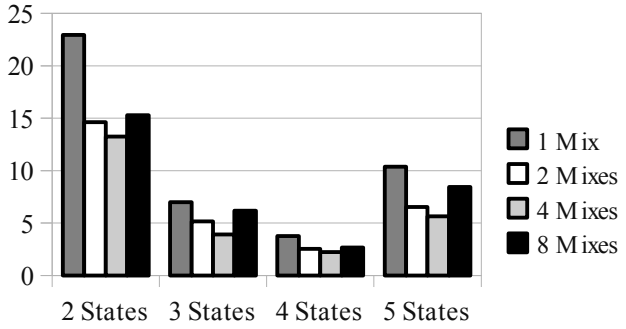

Figure 1: Inconsistency between labels, expressed as the mean rank of systems with MFCC front ends

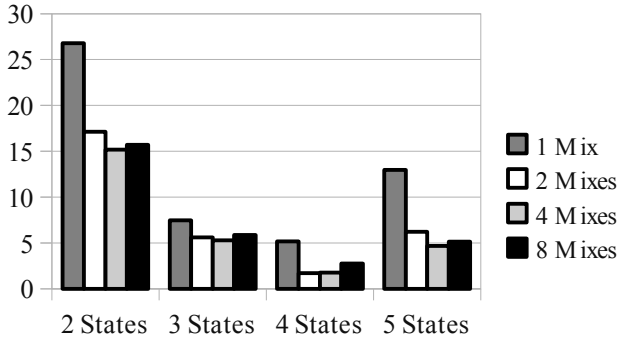

Figure 2: Inconsistency between labels of systems with LP Cepstrum front ends

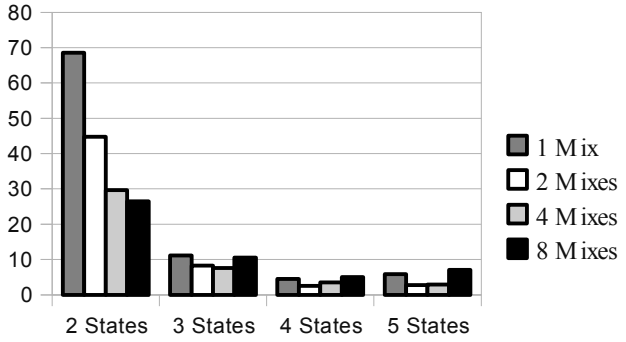

Figure 3: Inconsistency between labels of systems with ADV front ends

Each of Figures 1 to 3 was derived from 16 different systems, giving 120 (i.e. (16 * 15) / 2) different system pairs, and the maximum possible rank of 105. Thus a rank of 105 would mean that the respective ASPA system *always* had a worse discrepancy than any of the others.

It is clear from these results that, regardless of front-end, 4 states per phoneme always gives the best performance (or very close to it). The optimal number of mixtures is either 2 or 4 depending on the front-end.

The results for all front ends, for 4 states per phoneme and 2 and 4 mixture components, are shown together in Figure 4.

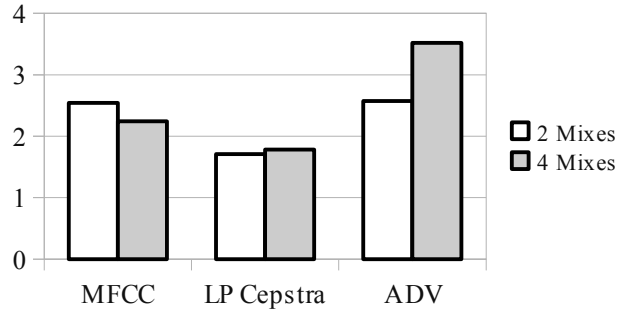

Figure 4: Inconsistency between labels of the best systems for each front end (4 states and 2 or 4 mixtures)

4.2. Parametric Bayesian Results

The figures below show system performance vs. front end, number of Gaussians per state, and number of states per phoneme, as quantified by the equation:

$$\sigma(aligner) = K_{\sigma}(aligner)\,\sigma_{global}$$

as derived from the parametric Bayesian models. This reflects each alignment system's overall inconsistency: it is the geometric mean of σ_{ab} for a system. The panels show the MFCC-based aligners (Figure 5), LPC-based aligners (Figure 6) and ADV-based aligners (Figure 7). The vertical axis represents the standard deviation in seconds.

Note that the ADV plot in Figure 7, below, has a very different structure to those of the "traditional" MFCC and LP Cepstrum front-ends.

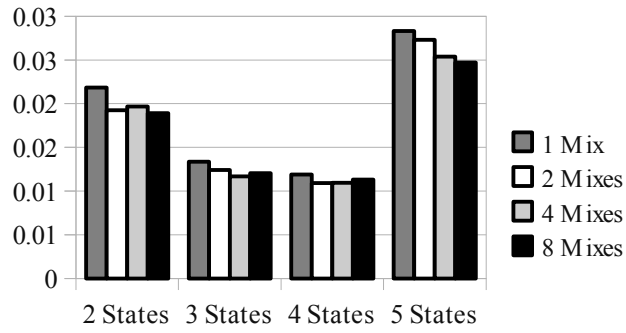

Figure 5: Alignment discrepancy standard deviations (in seconds) for systems with MFCC front ends

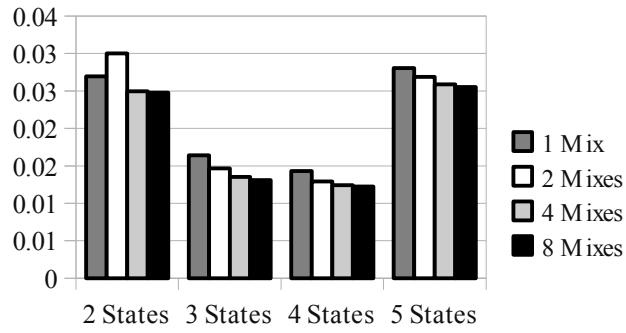

Figure 6: Alignment discrepancy standard deviations (in seconds) for systems with LP Cepstrum front ends

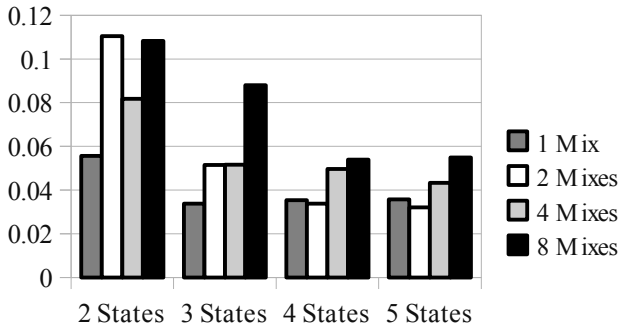

Figure 7: Alignment discrepancy standard deviations (in seconds) for systems with ADV front ends

4.3. Discussion

As can be seen in Figures 5 to 7, the Parametric Bayesian analysis gives very much the same qualitative results as the pairwise-variance-based analysis for the MFCC and LPC-based analyses: the number of states is the critical factor. Three or four states are much better than one or five, as can be seen by the lower values in the central two columns. Four states are slightly preferred to three. For both analyses, the number of Gaussians per state makes little difference to the performance as long as there is more than one.

However, the ADV results are rather more difficult to reconcile in the case of single mixtures: parametric Bayesian results are surprisingly good in this case, both relative to those calculated by pairwise variance, and relative to the parametric Bayesian results for larger numbers of mixtures.

Overall, MFCC performance is better than LPC by approximately 15% and in this analysis, both LPC and MFCC are dramatically better than the ADV-based systems (with standard deviations roughly three times larger). As most of the windows involved in the ADV calculations are larger than the 25 ms windows used in LPC and MFCC, it is not too surprising that the ADV captures timing less precisely and therefore gives less precise labels.

As regards the diagnostic value of these analysis methods, Figures 1 to 4 show that there is a fundamental difference between the ADV processor and the others. This difference manifests itself as a bias towards models with fewer mixture components but more states. This suggests that the ADV processor may be responding to temporal features on a different scale from conventional methods.

The ADV front end may be better at resolving sub-phonemic details, and that would also explain its apparent variability in the exact timing of its boundaries: it may be responding to sub-phonemic details in the acoustic signal, which are invisible in the other processors.

5. Conclusions

Both the "pairwise variance" and "parametric Bayesian" methods show that the LP Cepstrum front end is roughly comparable to MFCCs, but the current implementation of Auditory Description Vectors, at least for the specific phoneme alignment/labelling task described here, is less accurate. For this task, the pairwise variance analysis suggests the optimal HMM configuration uses 4 states per phoneme model and 2 mixture components with the LP Cepstrum front end. The parametric Bayesian analysis prefers MFCCs, with 4 states and 4 mixtures, but the differences between the two systems are small for both analysis methods.

Thus it can be inferred that MFCCs and LP Cepstra are broadly similar, the best number of states per phoneme is 4, with 2 to 4 Gaussians per mixture. This is slightly more states than are normally used in ASR (i.e. 3), and significantly fewer Gaussians than most speaker-independent ASR systems (8 or more). The most likely reason for this small number of mixtures is that the results presented here were derived from examples of a single dialect. A more diverse range of speakers would be expected to require more Gaussians per state.

More importantly, we have demonstrated two related methods for quantifying the reliability of labelling systems from the discrepancies between labels from a cohort of automatic systems, without any manually created reference labels. This facilitates rapid progress in optimising the design of any alignment system, and frees that process from the need for extensive manual intervention.

References

Baghai-Ravary, L., 1995, "Multi-dimensional adaptive signal processing, with application to speech recognition, speech coding and image compression". University of Sheffield PhD Thesis.

Beet, S. W., and Gransden, I. R., 1992, "Interfacing an auditory model to a parametric speech recogniser". Proc. IoA, Vol. 14, Pt. 6, pp 321-328.

Hutchinson, W., and Knopoff, L., 1978, "The acoustic component of Western consonance". Interface 7, 1-29.

Kochanski, G., et al., 2005, "Loudness predicts prominence; fundamental frequency lends little". J. Acoustical Society of America, 11(2):1038-1054. 2005.

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

Kochanski, G. and Rosner, B. S., "Maximum Likelihood Solutions for The Law of Categorical Judgement (Corrected), submitted to Psychometrika, 2009

Lander, T., 1997, "CSLU labeling guide". Center for Spoken Language Understanding, Oregon Graduate Institute.

Moore, B. C. J., and Glasberg, B. R., 1983, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns". J. Acoustical Society of America, 74(3):750-3, September 1983.

Sebestyen, G. S., 1962, "Decision-making processes in pattern recognition". ACM Monograph Series, MacMillan, pp 40-47.

SoX, 2009, "SoX Sound eXchange manual". http://sox.sourceforge.net/sox.html

Young, S. J., et al., 2006, "The HTK Book (for HTK Version 3.4)". Cambridge University Engineering Department. http://htk.eng.cam.ac.uk/docs/docs.shtml