Published in final edited form as:
Speech Commun. 2009 April ; 51(4): 352–368. doi:10.1016/j.specom.2008.11.003.

Speaker-Independent Phoneme Alignment Using Transition-Dependent States

John-Paul Hosom
Center for Spoken Language Understanding, School of Science & Engineering, Oregon Health & Science University, 20000 NW Walker Road, Beaverton, OR 97006 USA, hosom@cslu.ogi.edu

Abstract

Determining the location of phonemes is important to a number of speech applications, including training of automatic speech recognition systems, building text-to-speech systems, and research on human speech processing. Agreement of humans on the location of phonemes is, on average, 93.78% within 20 msec on a variety of corpora, and 93.49% within 20 msec on the TIMIT corpus. We describe a baseline forced-alignment system and a proposed system with several modifications to this baseline. Modifications include the addition of energy-based features to the standard cepstral feature set, the use of probabilities of a state transition given an observation, and the computation of probabilities of distinctive phonetic features instead of phoneme-level probabilities. Performance of the baseline system on the test partition of the TIMIT corpus is 91.48% within 20 msec, and performance of the proposed system on this corpus is 93.36% within 20 msec. The results of the proposed system are a 22% relative reduction in error over the baseline system, and a 14% reduction in error over results from a non-HMM alignment system. This result of 93.36% agreement is the best known reported result on the TIMIT corpus.

Keywords

forced alignment; phoneme alignment; automatic phoneme alignment; hidden Markov models

1. Introduction

1.1 Definition of Automatic Phoneme Alignment

The phoneme is a unit of speech that, by definition, differentiates one word from another (Ladefoged, 1993, p. 26). One phoneme may contain a number of distinct acoustic events. For example, a stop phoneme may consist of a sequence of closure, burst, and aspiration events; or, a diphthong may transition from a back vowel to a front vowel. However, not all acoustic events are phonemically relevant. In American English, there is no phonemic difference between stops that are aspirated and unaspirated, because the meaning of a word does not change with the degree of stop aspiration. Phonemes, therefore, provide a description of the speech signal at a level of abstraction that is especially useful for word-level speech processing. A speech signal may be described not only by *what* phonemes it contains, but also *where* its phonemes are located. The issue addressed in this paper is determination of the location of phonemes in a speech signal, given the sequence of phonemes contained in that signal. This

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

task of time-aligning phonemes in a waveform is referred to as “phoneme alignment,” and in the case of computer-based phoneme alignment, is called “automatic phoneme alignment.” In this article, we focus on the task of speaker-independent phoneme alignment, in which speaker characteristics and identity may change from utterance to utterance.

1.2 Applications of Automatic Phoneme Alignment

Determining the location of known phonemes is important to a number of speech applications. Most current automatic speech recognition (ASR) systems use context-dependent phonemes as their basic unit of recognition. When developing an ASR system, “good initial estimates … are essential” when training Gaussian Mixture Model (GMM) parameters (Rabiner and Juang, 1993, p. 370). Because states are associated with phonemes, the mean vectors and covariance matrices of each state’s GMM should be initialized with values that reflect characteristics of the phoneme associated with that state, as opposed to using initial values that are obtained without reference to phoneme identities (e.g. a “flat start” initialization). This, in turn, implies the need for reasonable estimates of the location of each phoneme when setting initial ASR system parameters, prior to locally-optimizing training techniques such as expectation-maximization. Phoneme location information is also critical when building concatenative text-to-speech systems. Even if a unit boundary is placed at the approximate center of a phoneme, so that concatenation cost is not as sensitive to boundary placement, errors in boundary placement will still negatively affect the duration and intersegmental timing of the synthesized speech.

Knowledge of phoneme boundaries is also necessary in some cases of health-related research on human speech processing. Studies that measure or modify speech intelligibility, such as investigations of age-related changes in the temporal processing of speech, or automatic improvement of the intelligibility of dysarthric speech, may require accurate determination of phoneme boundaries in order to modify speech stimuli in the correct regions of the signal (e.g. Gordon-Salant et al., 2006; Kain et al., 2007). One proposed diagnostic marker for Childhood Apraxia of Speech (CAS) (Shriberg et al., 2003a) requires accurate measurement of phoneme durations of known words. Another diagnostic marker for CAS and a proposed diagnostic marker for Alzheimer’s Disease both require estimation of which regions of the signal contain speech events and which regions contain pause events (Shriberg et al., 2003b; Singh et al., 2001). The location of speech and pause events can be determined, given an orthographic transcription of the speech signal, using a slight relaxation in the definition of automatic phoneme alignment, in which the phoneme sequence is determined from a pronunciation dictionary but the presence or absence of pauses (and their durations) is automatically determined.

In addition to these uses of phoneme alignment, there is a direct relationship between the most common method of automatic phoneme alignment, called “forced alignment,” and the Hidden Markov Model (HMM) framework used in most ASR systems. (This relationship is described in more detail in Section 2.2.) Therefore, improvements in the accuracy of forced alignment have the potential to translate into improvements in ASR word-recognition accuracy.

1.3 Measuring Phoneme Alignment

A primary difficulty in the task of phoneme alignment is that the boundary between two phonemes can be inherently subjective. Many types of boundaries are readily located by a human expert, based on identification of acoustic changes that are related to changes in the manner of articulation. As examples, the onset of voicing is characterized by the appearance of periodic glottal pulses in both the waveform and the spectrogram, and raising of the velum shunts airflow from the nasal tract to the oral tract and causes a sudden increase in energy, especially above 400 Hz. Other boundaries, such as the boundary between a /w/ and vowel, or between a vowel and an /l/, do not have clear transition points. (Phonemes are written here in TIMIT notation (Garofolo et al., 1990)). In these cases, because of coarticulation, the transition between phonemes may be a gradual change from one acoustic pattern to the next, with no precise boundary. Because of the lack of distinct physiological or acoustic events that signal a phoneme boundary in these cases, the location of the boundary is subjective (Svendsen and Kvale, 1990; Torkkola, 1988; Brugnara et al., 1993; Pellom, 1998; Ljolje et al, 1997). A convenient metric may be defined, such as the point of maximum slope in the second formant, but other criteria are just as valid. In continuous speech, boundaries may be difficult to reliably locate due to glottalization, extremely reduced vowels, or gradual decrease in energy before a pause. As a result, there is no “correct” answer to the phoneme alignment problem, because phoneme boundary placement is an inherently subjective task. Instead, we can measure the *agreement* between two alignments, such as the agreement between two humans, or the agreement between human and machine. Although precise evaluation of the quality of phonetic alignment is difficult, there is a general consensus that manual alignment is more accurate than automatic alignment (Ljolje et al., 1997; Cosi et al., 1991; Cox et al., 1998). Therefore, the goal of automatic phoneme alignment is not to achieve 100% accuracy, but to achieve agreement in boundary placement that is always as good as the best human-human agreement.

The agreement of automatic alignment with manual alignment is most often reported in terms of what percentage of the automatic-alignment boundaries are within a given time threshold of the manually-aligned boundaries. For example, Brugnara *et al.* (1993) report that for their system, 88.9% of the automatic boundaries are within 20 msec of the manual boundaries. This type of result will be reported here as a percent “agreement” within the given threshold; in this example, Brugnara’s system has 88.9% agreement within 20 msec. Results with a threshold of 20 msec will be reported when possible, as this threshold is commonly reported in other studies and allows a direct comparison between systems using a single value. Relative differences in the agreement between two systems will be reported using the terminology “reduction in error,” even though alternate terminology such as “increase in agreement” may be technically more correct. Manual alignment agreement is usually reported as inter-labeler agreement, with one set of manual alignments chosen as nominally “correct,” and the other set of alignments measured in relation to the first set.

The possibility arises that, because of human variability and machine consistency, automatic phoneme alignments have the potential to be better for some applications than manual phonetic alignments. However, due to a lack of evaluation methodology for this goal and current difficulties in accurate machine processing of speech, the focus of this paper is on achieving human-human levels of agreement using automatic phoneme alignment.

Although manual alignment is considered more accurate than automatic methods, manual alignment is too time consuming and expensive to be commonly employed for aligning large corpora. Manual alignment has been reported to take between 11 and 30 seconds per phoneme (Kvale, 1994; Leung and Zue, 1984), whereas automatic segmentation can be accomplished in real-time or faster. Because of the time requirements and inherent subjectivity of manual phoneme alignment, “there is a need for a fast, inexpensive, and accurate means of obtaining time-aligned phonetic labeling of arbitrary speech” (Wightman and Talkin, 1997).

1.4 Organization of Article

The remainder of this article is organized as follows: First, in the Background (Section 2), prior work on both manual and automatic phoneme alignment is discussed. Then, Section 3 describes the baseline alignment system, which uses a Hidden Markov Model/Artificial Neural Network hybrid (HMM/ANN) system (e.g. Bourlard et al., 1992; Hosom et al, 1998) for forced alignment. The proposed method, which incorporates energy-based features and transition-dependent states into the HMM/ANN framework, is then presented in Section 4. Sections 5 and 6 discuss the methods used in evaluation and the results from both the baseline and proposed technique. Finally, in the Conclusion (Section 7), results are summarized, public access to the baseline and proposed systems is described, and future work is discussed.

2. Background

2.1 Manual Phoneme Alignment

As noted above, the most accurate method of creating time-aligned phonetic labels is to employ an expert human labeler. This person typically generates phonetic alignments using a software tool that displays the speech waveform, spectrogram, phonetic labels, and possibly other acoustic information. The labeler aligns phonetic labels with the speech signal by listening to segments of the waveform and by using knowledge of the relationship between the waveform, its spectrogram, and its phonetic content. As a result, training in phonetics and spectrogram reading is required to produce acceptable label alignments, and manual alignment is a resource-intensive method.

Cosi et al. (1991) reported on the manual alignment of 10 continuous-speech Italian sentences recorded at 16 kHz and aligned by three people. They found a mean deviation of 6 msec, about 55% agreement within 5 msec, and 93.5% agreement within 20 msec. Ljolje et al. (1997) reported on the manual alignment for Italian utterances from two human transcribers, and found 80.0% agreement within 10 msec, 92.9% agreement within 20 msec, and 96.8% agreement within 30 msec. These results correspond well with those reported by Cosi. Wesenick and Kipp (1996) evaluated the manual alignment of German sentences by three transcribers. They found average agreement levels of 63% within 0 msec (perfect correspondence), 73% within 5 msec, 87% within 10 msec, and 96% within 20 msec. The transcribers in this study were all graduate students in phonetics, and all had received an intensive training session. As part of this training, a number of conventions were established to ensure consistent labeling . One such rule was to always set a segmentation boundary where the values of the speech signal changed from negative to positive (personal communication, Sep. 16, 1999). Not surprisingly, these results represent the best reported performance of human agreement on the task of phonetic alignment. Leung and Zue (1984) evaluated 5 American English sentences from the Harvard list of phonetically-balanced sentences, aligned by two people. Manual alignment required about 30 seconds per phoneme, and they reported approximately 80% agreement within 10 msec, 87% agreement within 15 msec, and 93% agreement within 20 msec. Cole et al. (1994) reported on inter-labeler agreement on telephone-channel speech for four languages, as labeled by both native and non-native speakers. For American English aligned by two transcribers (native speakers), they reported 79% agreement within 10 msec, which is marginally lower than the value reported by Leung. For German speech, they found 63% agreement within 5 msec and 79% within 10 msec when comparing two native-speaker labelers, and 69% agreement within 5 msec and 81% agreement within 10 msec when comparing a native-speaker labeler and a non-native-speaker labeler. One point of interest is that although Cosi, Ljolje, and Leung performed their evaluations on 16-kHz microphone speech and Cole et al. performed their evaluation on 8-kHz telephone-band speech, the results are quite comparable. In addition, the results from Leung on English speech and the results from Ljolje on Italian speech are nearly identical.

As none of the above evaluations were performed on the commonly-used TIMIT corpus of American English speech (Garofolo et al., 1990), we manually aligned 50 sentences from the test partition of TIMIT (1812 phoneme boundaries). Alignment was performed by the author, who had several years experience with reading spectrograms and performing phonetic alignment. We took the phoneme sequence as specified in the TIMIT phoneme-label files, but removed all timing information prior to manual labeling. For evaluation, we (a) removed the glottalization symbol /q/ by merging this symbol with surrounding voiced sounds and (b) did not evaluate boundaries between stop closures and silence (as any such boundary is placed arbitrarily). We found 81.73% agreement with the standard TIMIT alignments with a threshold of 10 msec, 93.49% agreement within 20 msec, and 96.91% agreement within 30 msec. These results correspond well with the results reported by Cosi, Ljolje, Leung, and Cole.

In summary, there is fairly consistent agreement among humans labelers for continuous speech, even across language and channel conditions. Including the results from Wesenick and Kipp, there is an average agreement of 93.78% within 20 msec (standard deviation 1.27%). Excluding the result of 96% from Wesenick and Kipp as an outlier due to a more precise set of labeling conventions, there is an average agreement of 93.22% within 20 msec (standard deviation 0.32%).

2.2 Prior Work on Automatic Phoneme Alignment

2.2.1 Overview of Alignment Systems—The most common method for automatic phoneme alignment is called "forced alignment." In this method, recognition of the speech signal is performed using an HMM, with the search path constrained to be the known sequence of phonemes. Because the Viterbi search can yield the locations of phoneme-based states as well as the state identities, phonetic alignment can be obtained by constraining the search to the known phoneme sequence. (These systems are called "forced alignment" systems because the alignment is obtained by forcing the recognition result to be the proposed phonetic sequence.) In general, there is a strong link between automatic speech recognition and forced-alignment techniques, in that many of the same general processes can often be used for both tasks. Dynamic Time Warping (DTW) is also used for automatic phoneme alignment (e.g. Wagner, 1981; Sevendsen and Soong, 1987; Gong and Haton, 1993; Campbell, 1996; Malfrère, Deroo, and Dutoit, 1998), but DTW tends to be applied to the task of speaker-dependent alignment.

Of 33 automatic alignment systems reported in the literature, 42% (14 systems) used HMM or HMM/ANN systems to obtain the alignments using forced alignment, and another 24% (8 systems) used DTW for speaker-dependent alignment. The remaining third (11 systems) employed a wide variety of approaches.

The most notable of the non-HMM and non-DTW systems was a system based on discriminative learning, in which phoneme boundaries were ranked "according to their quality" (Keshet et al., 2005). In particular, this system learned a set of functions that mapped from *both* observations and phoneme identities to phoneme start times. (In contrast, in the HMM framework, there is a mapping is from observations to phoneme (state) sequences, and the alignment is deduced from the phoneme assignments at each frame.) The learning procedure was similar to a Support Vector Machine, but instead of a binary classification, the learning process used a cost function for "assessing the quality of alignments." This system had results on the TIMIT test set of 92.3% within 20 msec.

2.2.2 Review of Forced Alignment Systems—Forced alignment is the dominant technique in automatic phoneme alignment. Rapp (1995) noted that because "the task of phoneme alignment can be considered as simplified speech recognition, it is natural to adopt a successful paradigm of ASR, namely HMMs, for alignment." In addition, forced-alignment results are generally superior to results from other methods. Here we review a few speaker-independent forced-alignment systems.

Wightman and Talkin (1997) developed a forced-alignment system called "the Aligner," with the acoustic model training and Viterbi search implemented using the HTK Toolkit (Woodland et al., 1995). The Aligner used a 10-msec frame rate, context-independent monophones, and a mixture of five Gaussians to estimate the state observation probabilities. Non-speech sounds, such as breath noise and lip smacks, were collapsed into a single “silence” model. The system was trained on stop closures separately from stop bursts, whereas other HMM systems often train the stop closure and the stop burst as one three-state phonetic unit. The Aligner was trained using the TIMIT labels for an initial segmentation. In evaluation of their system, they did not use the TIMIT phoneme sequence directly, but they first mapped the words to canonical dictionary pronunciations, then performed forced alignment, and finally mapped the forced-alignment phonemes to the TIMIT phoneme sequence. This indirect measurement allowed them to evaluate phonetic boundary alignments while performing forced alignment from only word-level information. Performance on the TIMIT test set using this metric was approximately 80% agreement within 20 msec.

Brugnara et al. (1993) developed an HMM-based forced-alignment system that used spectral variation features in addition to the standard cepstral-domain features for computing state occupation probabilities. The incorporation of these additional features resulted in a 2% relative reduction in error. They also adjusted the phonetic alignments after the Viterbi search, based on the values of the spectral variation features, but found no improvement in performance. They evaluated this system on the entire test partition of the TIMIT database, and reported 74.6% agreement within 10 msec, 88.8% agreement within 20 msec, and 94.1% agreement within 30 msec.

Pellom (1998) developed an HMM for forced alignment with a variety of speech-enhancement algorithms. This system used a 5-msec frame rate, 5-state HMMs, gender-dependent models, and a 16-component Gaussian Mixture Model at each state. When phoneme-level transcriptions were not available, the system generated pronunciations using the CMU dictionary and word-juncture modeling. The system was trained and evaluated on TIMIT data that had been downsampled to 8 kHz, which resulted in 86.2% agreement within 20 msec.

Ljolje and Riley (1991) built an HMM system, with three states per phoneme, that used different types of phonetic models depending on the availability of training data. If enough data were available for a given phoneme in its left and right contexts, then a complete triphone model was used, although the left and right contexts were clusters of similar phonemes instead of individual phonemes. If sufficient data were not available for a full triphone model, then a “quasi-triphone” model was attempted; this quasi-triphone model had the left state dependent on the left context, the middle state context independent, and the right state dependent on the right context. If sufficient data were not available for the “quasi-triphone” model, then left-context dependent and right-context dependent models were both attempted. If sufficient data were still not available, then context-independent phoneme models were used. The HMM used full-covariance Gaussian probability density functions to estimate the observation probabilities, a Gamma-distribution duration model, and a 10-msec frame rate. The models were trained and evaluated on the TIMIT database. Two types of models were trained: those based on the manual alignments in the TIMIT database, and those based on a mixture of manual alignments and Viterbi re-estimation of the alignments. In either case, they found 80% agreement within 15 msec.

In summary, reported forced-alignment systems employ numerous modifications to the standard HMM training procedure, but in all cases the basic HMM process remains the same. Direct comparison of the results from these systems is not possible, because even in four cases where the systems were evaluated on the same corpus (TIMIT), there were minor implementation differences that prevent a one-to-one benchmark comparison. In the case of Pellom’s system, the TIMIT corpus was down-sampled to 8 kHz for training and evaluation, the frame rate was 5 msec, stop closures were merged with their succeeding plosives, and there was a total of 46 phonemes; in Brugnara’s system, training used 16 kHz data, the frame rate was 5 msec, stop closures were not merged with their succeeding plosives, and there was a total of 48 phonemes. Ljolje and Riley trained at 16 kHz with a 10-msec frame rate, merged stop closures with their bursts, and used a set of 47 phonemes. Wightman and Talkin trained at 16 kHz, used a 10-msec frame rate, did not merge stop closures, and used a set of 35 phonemes. If, however, we assume that the performance differences due to these variations are minimal (e.g., Wightman and Talkin claim “very similar” results for systems trained on 16 kHz and 8 kHz speech), we can generally conclude that performance of forced-alignment systems on the TIMIT database ranges from 80% to 88.9% agreement within 20 msec. Performance on other databases and languages tends to be similar but slightly lower, with agreement levels from 77% to 84% within 20 msec. Only Pellom (1998) and Wheatley et al. (1992) evaluated systems on telephone-band speech, and severe performance degradation was reported; even systems with the best possible noise compensation had no more than 76.8% agreement within 20 msec for landline telephone speech and 66.7% agreement within 20 msec for cellular-telephone speech.

2.3 Summary of Prior Work on Phoneme Alignment

Manual alignment is reported to have inter-labeler agreement between 92.9% (Ljolje et al., 1997) and 96% (Wesenick and Kipp, 1996) within 20 msec, with an average of 93.78% within 20 msec. Manual-alignment agreement on the TIMIT corpus is 93.49% within 20 msec. The inter-labeler agreement on TIMIT is consistently higher than automatic-alignment agreement across all thresholds and for all systems. Previously-reported HMM-based automatic alignment systems have maximum agreement of 88.9% within 20 msec (Brugnara et al., 1993). The best reported agreement for a non-HMM alignment system is 92.3% within 20 msec using a discriminative learning system (Keshet et al, 2005).

3. Baseline System

In order to evaluate the proposed technique, a baseline forced-alignment system was developed on the same data and using the same phoneme set as the proposed system (Section 4). This baseline system was an HMM/ANN hybrid (e.g. Bourlard et al., 1992; Hosom et al, 1998), which computes probability estimates of observations using an artificial neural network (ANN) instead of a Gaussian Mixture Model (GMM). The general framework presented here has been used previously on a variety of tasks including digit recognition (Hosom et al., 1998), children’s speech recognition (Shobaki et al., 2000), and recognition of Italian (Cosi and Hosom, 2000) and Vietnamese (Duc et al., 2003). The parameters used in the baseline forced-alignment system are summarized in Table 1.

3.1 Training Data and Features

The baseline system was trained on 3696 files (3.145 hours of speech) from the training partition of the TIMIT corpus (excluding “sa” files). The feature set consisted of features similar to Mel-Frequency Cepstral Coefficients, except that the Bark frequency scale was used instead of the mel scale. (While there isn’t much practical difference between the two frequency scales, “the traditional mel scale has in many technical fields been replaced by the Bark scale” (Huopaniemi and Karjalainen, 1997).) The steps of feature extraction were (a) pre-emphasis (with a factor of 0.97), (b) application of a 24-msec Hamming window, (c) computation of the power spectrum, (d) non-linear frequency warping of the power spectrum along the Bark scale using 40 filters, (e) conversion of power-spectrum values to the logarithm domain, (f) conversion to cepstral features using the inverse discrete cosine transformation of these frequency-warped log power spectrum values, and (g) exponential weighting by a factor of 0.6 to increase the weight of higher cepstral coefficients. Only the lowest 13 cepstral coefficients were selected in the final feature set. Features were computed with a 16-kHz sampling frequency and a 5-msec frame rate. The lowest cepstral coefficient was replaced with the log energy of the signal, computed with a 100-msec Hamming window, and normalized by the maximum and minimum energy values to be within the range -1.0 to 1.0.

A variant of cepstral mean subtraction (CMS) was used, in which the cepstral values subtracted from the feature set were not the mean cepstral values over the entire utterance, but the mean cepstral values over the 100-msec region of the signal with lowest energy. This technique, "Low-Energy Cepstral Mean Subtraction," may be more effective on very short utterances than on the TIMIT corpus, but was used here as a feature-processing method that is independent of utterance length.

The delta values of the 13 cepstral coefficients were included in the feature set, but acceleration coefficients were not included, as their impact tends to be minimal on the type of HMM/ANN system described here.

3.2 Phoneme Set and Context-Dependent Categories

The complete set of 61 TIMIT phoneme symbols was mapped to a set of 54 phonemes as follows. First, the sentence-beginning and sentence-ending pause symbols /h#/ were mapped to pause (/pau/). Epenthetic silence (/epi/) was also mapped to pause. The syllabic phonemes /em/, /en/, /eng/, and /el/ were mapped to their non-syllabic counterparts /m/, /n/, /ng/, and /l/, respectively. The glottal closure symbol /q/ was removed, as it is used in TIMIT sometimes to annotate a glottal stop consonant (e.g. dr1/fcjf0/sa1), sometimes to indicate glottalization (e.g. dr1/fcjf0/sx127), and other times to indicate an unusual acoustic event (e.g. dr1/fcjf0/si1027). If the glottal closure neighbored a voiced phoneme on one side and an unvoiced phoneme on the other side, the glottal closure was merged with the voiced phoneme. If the glottal closure was surrounded by two voiced phonemes, then the boundary of the two neighboring phonemes was placed at the mid-point of where the glottal closure occurred. If the glottal closure was surrounded by two unvoiced phonemes, it was mapped to a short neutral vowel, /ax/. Finally, short pauses with duration less than 20 msec were removed. If the short pause neighbored a voiced phoneme on one side and an unvoiced phoneme on the other side, the short pause was merged with the unvoiced phoneme. Otherwise, the boundary of the two neighboring phonemes was placed at the mid-point of where the short pause occurred.

In addition to this mapping, phonemes with significant acoustic variation were split into sub-phonetic units in order to better model the acoustic dynamics of these phonemes. The diphthongs /ay/, /oy/, /aw/, /ey/, and /ow/ were split into two parts, one for the first two-thirds of the phoneme and the other for the final third. The affricates /ch/ and /jh/ were also split into two parts, with the first 10-msec part corresponding to the burst and the second, remaining, part corresponding to the frication. These split diphthongs and affricates were mapped back to their corresponding whole diphthongs or affricates after alignment but before evaluation. These splits resulted in 61 phonetic and sub-phonetic units for classification.

The baseline alignment system used 451 states to represent these 61 phonetic and sub-phonetic units. Those phonemes that are heavily influenced by coarticulation (vowels, semivowels, and /h/) were modeled using three states per phoneme, the first and last of which were context-dependent. Liquids and glides were modeled using two context-dependent states per phoneme, because they tend to be influenced by coarticulation but are shorter in duration than vowels. Nasals, stops, flaps, fricatives, affricates, and pauses were modeled using one context-independent state per phoneme. To provide sufficient data per state for training, a context-dependent phoneme's initial state was clustered according to its left context, and a context-dependent phoneme's final state was clustered according to its right context. The middle state of a three-state phoneme was always modeled independently of its left or right context. (This method of defining context-dependent states is equivalent to the "quasi-triphone" model of Ljolje and Riley (1991).) The left or right contexts consisted of ten broad-phonetic classes, namely silence, front vowel, mid vowel, back vowel, retroflex, lateral, labial, dental, alveolar, and dorsal. For example, the phoneme /iy/ in the context /s iy p/ was represented using three states, /iy/ in the context of a preceding alveolar, the context-independent central region of the /iy/, and /iy/ in the context of a subsequent bilabial. This clustering resulted in 680 states. States with less than 32 training examples were then tied to states with a greater number of examples, resulting in the final set of 451 states.

3.3 Training of ANN for HMM/ANN Hybrid

The probability of an observation given a state was estimated using an 3-layer ANN trained on up to 32,000 examples per state. The ANN had as input features a context window of five frames, with frames at -60, -30, 0, 30, and 60 msec relative to the center frame. The network thus had an input layer of 130 nodes (13 features per frame and 5 frames), a hidden layer of 300 nodes, and an output layer of 451 nodes. Training was performed for 45 iterations, and the weights from the 45th iteration were used for the final system parameters. The learning rate at each iteration was 0.05/((0.2*i*)+1.0) where *i* is the iteration number (from 1 to 45), thus decreasing from 0.0417 at the first iteration to 0.0066 at the final iteration.

Given assumptions about the quantity of training data, correct number of hidden nodes, and other factors, the output of an ANN can be considered to approximate the probability of each class given the input observation, $p(j \mid \mathbf{o}_t)$, where j is the class (or state of the HMM) and $\mathbf{o}_t$ is the observation (or set of features) at time t (e.g. Richard and Lippmann, 1991). For a Hidden Markov Model, the probability of the observation given the state is required, $p(\mathbf{o}_t \mid j)$. Using Bayes' rule, it is possible to convert $p(j \mid \mathbf{o}_t)$ to a scaled representation of $p(\mathbf{o}_t \mid j)$ by dividing by the *a priori* probability of the state, $p(j)$; the scaling factor $p(\mathbf{o}_t)$ does not need to be computed because it is constant for all states and thus does not impact the maximization operation during the Viterbi search. However, $p(j)$ can be quite small, and division by a small number can yield large errors if that small number is not accurate. One solution is to train each category using the same number of examples, thereby removing the effect of $p(j)$ during training. This is not always possible in practice, as some phonetic categories have many more examples than other categories. Therefore, we apply a “negative penalty” modification to the training procedure (Wei and van Vuuren, 1998) in order to remove the effect of $p(j)$ during training while still using a different number of examples per class. This negative penalty modification is accomplished by weighting the neural network cost function for infrequently-occurring categories, based on the statistics of the training data. The outputs of the ANN are therefore considered scaled estimates of $p(\mathbf{o}_t \mid j)$, and are used directly by the Viterbi search during decoding.

3.4 Duration Modeling

The standard HMM has, because of its transition probabilities, an implicit duration model that models the probability of staying in any state for exactly t frames decaying exponentially as a function of t (Rabiner and Juang, 1993, p. 358). It has been noted that this duration model does not fit well with observed durations of phonemes, which have durations that are better modeled using a Gamma probability density function (Levinson, 1986). Semi-Markov models (SMMs) have been proposed to address this issue (e.g. Levinson, 1986), but SMMs are computationally much more expensive than standard HMMs. The duration model of the baseline system provides an approximation to a Gamma distribution with the same computational cost as a standard HMM. In this duration model, state durations less than $D_{min}(j)$ frames have an exponentially decreasing probability as the duration of state j decreases, and state durations greater than $D_{max}(j)$ frames have an exponentially decreasing probability as the duration of state j increases, where $D_{min}(j)$ and $D_{max}(j)$ are state-dependent duration parameters that represent target minimum and maximum durations, respectively. The (non-normalized) probability of being in state j for d frames, when d is less than $D_{min}(j)$, is computed as

$x_1^{(Dmin(j)-d)}$, where (for simplicity) x_1 is a state-independent value of 0.0015. The (non-normalized) probability of being in a state *j* for *d* frames, when *d* is greater than $D_{max}(j)$, is computed as $x_2^{(d-Dmax(j))}$, where x_2 is a state-independent value of 0.368. These probability functions and values of x_1 (0.0015) and x_2 (0.368) have been empirically determined on other datasets to provide reasonable results. The D_{min} value for state *j* is set to the duration of the 2nd percentile of all durations for this state in the labeled training data. The D_{max} value for state *j* is set to the maximum duration for this state in the labeled training data. These state-duration probabilities can be applied during the Viterbi search and used in lieu of standard state-transition probabilities.

4. Proposed System

The proposed system implements three modifications to the baseline system: (1) The feature set includes, in addition to the baseline system's cepstral features and normalized log energy (computed with a 100-msec window), four additional energy-based feature streams; (2) The system uses, in addition to probabilities of each phoneme-based state given an observation, probabilities of a state transition given that observation; and (3) Instead of computing context-dependent phoneme probabilities directly, the system computes the probabilities of distinctive phonetic features. The probabilities of these features are then combined to obtain phoneme probabilities. Each of these three modifications is described in more detail below.

4.1 Additional Features

The additional features in the proposed system were designed to be robust and provide some degree of information complementary to the cepstral feature set. These four feature streams, described in more detail in the following paragraphs, included: (1) an intensity-discrimination feature for the entire signal and for seven frequency bands, (2) the time derivatives of these intensity-discrimination features, (3) a relative-energy based burst-detection feature (Hosom and Cole, 2000), and (4) normalized log-scale energy computed with a 40-msec Hamming window, to focus on energy changes that are more rapid than can be measured with the 100-msec energy window used in Section 3.1.

4.1.1. Intensity Discrimination—The intensity-discrimination feature is motivated by perceptual studies of the smallest change in acoustic intensity that is detectable by humans; these studies have been summarized by Moore (1997, pp. 63-65) and can be generally modeled using the following equation:

$$\Delta L_t = 10 \log \left(\frac{\Delta E_t}{E_t} + 1 \right) \tag{1}$$

where ΔL_t is the measurement of intensity discrimination at time *t*, ΔE_t is the change in intensity of the signal at time *t*, and E_t is the reference intensity of the signal at time *t*. Moore defines intensity as the sound power transmitted through a given area of the sound field, although it can be used to described "any quantity relating to the amount of sound, such as power or energy." (Moore, 1997, p. 361). We compute ΔE_t and E_t using Hamming-windowed energy, with different window lengths to identify certain types of energy changes and reference intensities. For identifying phoneme-level intensity changes, a window length of 40 msec is used for computing ΔE_t, and a window length of 250 msec is used for computing E_t. The value of 40 msec was chosen to correspond to the minimum duration of a speech segment required for assigning phonetic quality (Greenberg, 1996). The value of 250 msec was chosen to correspond to the approximate duration of a syllable. To obtain smooth delta values, ΔE_t is computed using Furui's equation for dynamic features (Furui, 1986),

$$\Delta E_t = \frac{\sum_{\theta=1}^{\Theta} \theta \cdot (E_{t+\theta} - E_{t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2} \tag{2}$$

with a Θ value corresponding to 20 msec (4 frames). The function ΔL_t is maximum when the *relative* change in energy is greatest, and the most negative when the relative change in energy is most negative. When the energy of the 40-msec signal does not change relative to the surrounding 4 frames, ΔL_t is zero. In addition to computing ΔL_t for the entire signal, we compute ΔL_t values for seven frequency bands spanning 210 Hz to 4000 Hz, each one bark in width: 210–430 Hz, 430–710 Hz, 710–1060 Hz, 1060–1530 Hz, 1530–2180 Hz, 2180–3060 Hz, and 3060 Hz–4000 Hz. Intensity discrimination in these frequency bands targets phoneme-related energy changes at relevant regions of the spectrum. While this intensity discrimination is conceptually similar to RASTA processing (Hermansky and Morgan, 1994), (a) intensity discrimination is easier to implement, given that the RASTA filter coefficients must be properly initialized, and (b) it operates on spectral energies instead of cepstral coefficients.

4.1.2. Time Derivatives of Intensity Discrimination—The time derivatives of these intensity-discrimination signals were considered to be potentially useful, especially with maximum and minimum values of ΔL_t identifying the instants of maximum and minimum energy change in the signal, and thus the time derivative $\Delta L_t'$ identifying maximum and minimum energy change with values of zero. The time derivatives were also computed using Furui's equation for dynamic features (Furui, 1986), again with a Θ value of 20 msec (4 frames).

4.1.3. Burst-Detection Feature—Stop bursts were identified using another relative-energy based feature (Hosom and Cole, 2000). Burst-related impulses (bursts) occur at the instant of release of stop phonemes, and can be described as a sudden impulse-like increase in energy due to release of air from the mouth. Each burst is produced by a closure of the oral cavity in order to create an increase in internal air pressure, followed by a sudden release of the constriction, which causes an abrupt increase in energy of the signal. Because of this process, bursts are characterized by at least 15 msec of low energy (during the closure), which is followed by a sudden increase in energy (at the instant of release), which is followed by a gradual decline in energy (during the release). Furthermore, the radiation characteristic of sound emanating from the mouth causes the burst at the instant of release to take on the qualities of an impulse, with a relatively flat spectrum and short duration. The spectral envelope of the burst is shaped to some degree according to the place of constriction. Therefore, (a) there must be a relative increase in energy at the instant of release, (b) the increase in energy must occur over most frequency bands, and (c) the burst must have certain spectral properties that distinguish it from environmental noise (such as clicks). The first two criteria can be detected by using Moore's measure of intensity discrimination on several frequency bands to estimate relative changes in energy (Equation 1), and then combining this frequency-band information (scaled to the range 0 to 1 and treated as the probability of a sudden increase in energy) into a single burst-detection feature using Bayes' rule. In later processing, this burst-detection feature can be used with other (e.g. cepstral) features as input to a classifier to incorporate spectral properties into the burst detection process. In this application, the window size of E_t was 22 msec, the window size of ΔE_t was 24 msec, the value of Θ was 10 msec (2 frames), and relative-energy information from seven frequency bands was computed and then combined (with frequency bands 210–430 Hz, 430–710 Hz, 710–1060 Hz, 1060–1530 Hz, 1530–2180 Hz, 2180–3060 Hz, and 3060–4000 Hz) using Bayes' Rule.

4.2 Observation-Dependent Transition Probabilities

In a standard Hidden Markov Model, the probability of an observation vector sequence, **O**, and state sequence, *Q*, is given by

$$P(\mathbf{O},Q)=P(\mathbf{O}|Q)\;\cdot\;P(Q) \tag{3}$$

which, with the assumptions of a first-order Markov process and independence of individual observations at each time *t*, $\mathbf{o}_t$, can be expanded as follows:

$$P(\mathbf{O},Q)=\prod_{t=1}^{T}p(\mathbf{o}_t|q_t)\;\cdot\;p(q_0)\;\cdot\;\prod_{t=1}^{T}p(q_t|q_{t-1}) \tag{4}$$

where *T* is the maximum time value and q_0 is the initial state. We can use Bayes' Rule, e.g. $P(A|B) = P(B|A) \cdot P(A)$ / P(B), to re-write the first term of this equation, yielding:

$$P(\mathbf{O},Q)=\prod_{t=1}^{T}\frac{p(q_t|\mathbf{o}_t)\,p(\mathbf{o}_t)}{p(q_t)}\;\cdot\;p(q_0)\;\cdot\;\prod_{t=1}^{T}p(q_t|q_{t-1}) \tag{5}$$

In the proposed system, each state q_t is linked with an additional state x_t. This state x_t indicates whether or not there is a phoneme transition at time *t*. If there is no phoneme transition at time *t*, then x_t is a state labeled "same," without regard to the identity of the phoneme. If there is a phoneme transition at time *t*, then x_t is labeled with both the phoneme at *t*-1 and the phoneme at time *t*. The association between states q_t and x_t is illustrated in Figure 1.

As an example using context-dependent states, the phoneme /iy/ in the context /s iy p/ would be represented using three states, /s<iy/, /iy/, and /iy>p/. At the transition between /s<iy/ and / iy/, the *x* state is "same", because there is no phoneme transition. At the transition between the /s/ phoneme's /s>iy/ state and the /iy/ phoneme's /iy<s/ state, however, the *x* state is the transition /s→iy/. Therefore, for the majority of states and state transitions, *x* is "same". The value of *x* is only phoneme-specific when there is a phoneme transition.

In the proposed system, we model the probability of an observation sequence and both state sequences as the combination of two HMMs:

$$\begin{aligned}P(\mathbf{O},Q,X)\quad=\quad&\prod_{t=1}^{T}\frac{p(q_t|\mathbf{o}_t)\,p(\mathbf{o}_t)}{p(q_t)}\;\cdot\;p(q_0)\;\cdot\;\prod_{t=1}^{T}p(q_t|q_{t-1})\,\cdot\\&\prod_{t=1}^{T}\frac{p(x_t|\mathbf{o}_t)\,p(\mathbf{o}_t)}{p(x_t)}\;\cdot\;p(x_0)\;\cdot\;\prod_{t=1}^{T}p(x_t|x_{t-1})\end{aligned} \tag{6}$$

As the transition probabilities are generally considered to have a minimal impact on system performance, and because of the increased complexity of keeping track of two state sequences in the Viterbi search, we have simplified the model by assuming that $p(x_0)$ and $p(x_t \mid x_{t-1})$ have negligible impact on final results and can be factored out. (As Huang, Acero, and Hon (2001) state, "In practice, duration models offer only modest improvement for speaker-independent continuous speech recognition. Many systems even eliminate the transition probability completely because the output probabilities are so dominant." (p. 408).)

We can then separately model $p(q_t \mid \mathbf{o}_t)$, the probability of a standard HMM state q_t given an observation $\mathbf{o}_t$, and $p(x_t \mid \mathbf{o}_t)$, the probability of a phoneme transition at time t given the observation $\mathbf{o}_t$. The method of obtaining estimates of $p(q_t \mid \mathbf{o}_t)$ and $p(x_t \mid \mathbf{o}_t)$ will be described in Section 4.3. With those estimates, standard forced alignment can be performed using a Viterbi search. The Viterbi search is simply modified to incorporate the observation-dependent probability of a phoneme transition at time t when making state transitions. If the states at t and t-1 belong to the same phoneme, then the phoneme transition is "same," and if the states belong to different phonemes, then the phoneme transition is specified from the state identities. Note that either $p(q_t \mid \mathbf{o}_t)$ or $p(x_t \mid \mathbf{o}_t)$ can be used individually to perform recognition or alignment, but that states q_t and x_t have been defined in such a way as to provide the most information where the probability estimate of the other state is least informative.

4.3 Use of Distinctive Phonetic Features

In typical HMMs, the basic unit of classification is the phoneme, which is usually divided into a number of sequential context-dependent states. In the proposed system, we split each context-dependent state into three (time-synchronous) parts that measure aspects of that phoneme's speech production properties. Specifically, the aspects that we considered are the distinctive phonetic features Manner (manner of articulation), Place (tongue position), and Height (height of tongue body). The probabilities of these aspects were estimated separately and then combined to arrive at a (context-dependent) phoneme-level probability. The context dependency was limited to the place of articulation, because this aspect has the greatest impact on non-local characteristics of the speech signal. As an example of the use of distinctive phonetic features, a /d/ in the context of a following /aa/ was modeled by (1) a voiced stop, (2) an alveolar tongue position in the context of a following back tongue position, and (3) a maximum tongue height. The use of distinctive features in the current work was primarily motivated by a desire to maximize the amount of training data per output node of the classifier, especially for the phoneme-transition categories, although other motivations (such as potentially modeling asynchrony between articulatory gestures) are still important factors.

For estimation of context-dependent phoneme probabilities $p(q_t \mid \mathbf{o}_t)$, we employed three parallel ANN classifiers, each trained to estimate the probability of a different distinctive phonetic feature, namely Manner, Place, or Height. Assuming conditional independence between the probabilities of these categories, we combined the distinctive-feature classification outputs for a single observation, using Bayes' rule, to arrive at phoneme-level probabilities. The three distinctive features and their values are specified in Table 2. (The exception to the use of Bayes' rule was for the "closure" category, which was common to all three features, and so we assumed complete dependence of the three closure probabilities and combined the probability values by averaging.) The features and their values were selected so that the maximum number of American English phonemes could be specified using the minimum number of feature values. There are two sets of phonemes not distinguished by the values specified in Table 2: the pair of vowels /aa/ and /ao/, and the pair of retroflex phonemes /er/ and /r/.

The probability of a distinctive phonetic feature f at time t, given an observation $\mathbf{o}_t$, e.g. $p(f_t \mid \mathbf{o}_t)$, was estimated using an ANN, with 158 input features (corresponding to the 130 cepstral features and 28 additional energy-based features) and 300 hidden nodes. The number of output categories of the ANN depended on the distinctive phonetic feature. Because coarticulation does not greatly affect the observed manner of articulation, the Manner classifier used context-independent categories. The Place classifier used context-dependent categories in order to better model the effects of coarticulation. The Height classifier used context-independent categories in order to simplify implementation details. For the Manner classifier, there were 10 categories (ANN outputs); for the Place classifier, there were 108 (context-dependent) categories; and for the Height classifier, there were 6 categories. Training was performed for 45 iterations, and the weights from the last iteration were used as the final classifier parameters. As in the baseline system, we applied a "negative penalty" modification to the training procedure (Wei and van Vuuren, 1998) in order to remove the effect of $p(f_t)$ during training, in effect estimating $p(f_t \mid \mathbf{o}_t)/p(f_t)$ or a scaled version of $p(\mathbf{o}_t \mid f_t)$.

The probabilities of phoneme transitions were estimated in a similar way, with three separate distinctive phonetic feature transitions combined using Bayes' rule to arrive at phoneme-level transition probabilities. In this case, the Manner-Transition classifier estimated the probability of each Manner transition (100 categories) as well as the probability that the observation was not at a phoneme boundary (1 category). The Place-Transition classifier estimated the probability of each Place transition (100 categories) as well as the probability that the observation was not at a phoneme boundary (1 category). The Height-Transition classifier had 36 categories for each Height transition, and 1 category for a non-boundary transition. The same 158 input features were used, and each classifier used 300 hidden nodes. Training was performed for 45 iterations, and the negative penalty modification was applied.

4.4 The Entire Proposed System

The entire system was constructed by (a) using the proposed energy-based features in addition to standard cepstral features as input to each ANN classifier, (b) applying Bayes' Rule to combine probabilities of distinctive phonetic features generated by the ANNs into probabilities of phonemes, and (c) modifying the Viterbi search to incorporate the probability $p(x_t \mid \mathbf{o}_t)$ with the standard transition and observation probabilities. The use of the six different ANN classifiers is illustrated in Figure 2 for the phoneme sequence /b aa t/, where both the /b/ and /t/ occupy one frame, and the /aa/ occupies two frames. The output of the Viterbi search contains the phoneme identities and boundaries.

5. Evaluation Method

Both the baseline and the proposed system were evaluated on the 1344 "si" and "sx" files in the testing partition of the TIMIT corpus. The input to the systems consisted of waveforms and their corresponding phoneme strings. A phoneme string was obtained by performing the phonetic mapping described in Section 3.2 on the existing time-aligned phoneme file (reducing the symbol set to 54 phonemes), and then removing the timing information. The output from the systems were time-aligned phoneme files using the same set of 54 phonemes.

The performance of a system was measured by computing the agreement between the TIMIT time-aligned phoneme files and the time-aligned phoneme files that were output by the system. Agreement was measured at thresholds of 5 msec, 10 msec, 15 msec, … 100 msec. In order to compare two systems using a single measurement, the commonly-used threshold of 20 msec was chosen.

6. Results and Discussion

6.1. Main Results

Results were computed on a total of 49,261 phoneme boundaries in the 1344 sentences (files). Results from the baseline system for each threshold are given in the second column of Table 3. In particular, the performance of this baseline system expressed as a single number is 91.48% within 20 msec. From these results, we conclude that the baseline system has better performance than any previously-published HMM-based speaker-independent alignment system that was evaluated on the TIMIT corpus. (Better performance on the TIMIT corpus was reported for a non-HMM alignment system developed by Keshet et al. (2005), with 92.3% agreement within 20 msec. The best previously-reported HMM-based results were from Brugnara et al. (1993), with 88.8% agreement within 20 msec on TIMIT.)

Results from the proposed system are given in the third column of Table 3. For this system, the agreement of 93.36% within 20 msec represents a 22.1% relative reduction in error over the baseline system, assuming a maximum agreement of 100%. Because agreement of 100% is impossible in practice, due to the inherently subjective nature of phonetic alignment, the actual reduction in error is larger, although not easily quantified. Using the McNemar significance test (Gillick and Cox, 1989), the difference between the two systems is significant ($p < 0.001$) at the 20-msec threshold. (It is acknowledged, however, that in general statistical significance is relatively easy to obtain with large datasets.) The proposed system demonstrates, at the 20-msec threshold, a relative 13.77% reduction in error over previously reported results on the TIMIT corpus (Keshet et al, 2005).

The agreement of manual alignments described in Section 2.1 are provided in column 4 of Table 3. These values were obtained on a subset of only 50 test sentences, and so they are not as precise as the results for the automatic systems, which were evaluated on 1344 sentences. The proposed system’s agreement of 93.36% within 20 msec is within a relative 2.0% of the manual agreement of 93.49% within 20 msec. At some thresholds, the automatic system has higher agreement levels than the manual alignments. Rather than concluding that the automatic system has better-than-human performance, we explain this result by considering two factors. First, the results of manual alignment were obtained from analysis of a subset of only 50 TIMIT sentences. Therefore, small differences between the automatic results and manual results may reflect a difference in evaluation data, and may not be statistically significant. Second, the automatic system has presumably learned specific characteristics of the canonical TIMIT alignments. In other words, the canonical TIMIT alignments may have been consistently labeled with slightly different (subjective) criteria than the current manual alignments. Therefore, we can expect the results of the proposed system to be somewhat higher when evaluated on test-set data containing canonical TIMIT alignments, as compared with evaluation on test-set data with alignments from different sources. We note that the accuracy of the manual alignment of TIMIT given here is comparable to other manual alignments at the 20-msec threshold (Section 2.1); at larger thresholds, these results on TIMIT are equal to or better than the results of Leung and Zue (on a corpus similar to TIMIT) (Leung & Zue, 1984). Therefore, the higher accuracy of the automatic system does not seem to be due to low accuracy of the manual alignments. The fifth column of Table 3 shows the results of evaluating the proposed system with the 50 manually-aligned TIMIT test sentences assumed to contain the correct boundary locations. It can be seen that the results are slightly better with the canonical alignments. This bias toward the canonical TIMIT alignments may explain, at least in part, why some of the automatic results (trained and evaluated on canonical TIMIT alignments) are better than manual results. In summary, the comparison of automatic with manual results should be noted with the conditions that (a) the comparison is not direct, but involves different amounts of data, and (b) the system was trained on canonical TIMIT alignments and may have learned specific, subjective characteristics of the canonical alignments.

It is also noted that the proposed system does not have better agreement than the baseline at all thresholds; at 5 msec, the baseline system has slightly better performance. At higher thresholds (e.g. 50 msec and higher), the difference between the two systems becomes negligible, with the proposed system having greater agreement in some cases, and the baseline system having greater agreement in other cases. Therefore, the proposed technique is most effective at thresholds between 10 and 50 msec, and does not have any advantage in correcting gross alignment errors. In general, in a speech-processing system in which large displacement of segments has a critical effect, the baseline system is about as effective as the proposed system, and either one can be used with equal effect. In a system in which small displacements are important, the proposed system provides an advantage.

6.2. Additional Tests and Results

Although it is difficult to separate the use of transition-dependent states from the use of distinctive phonetic features due to the way in which the proposed system was implemented, the energy-based feature set is independent of both distinctive phonetic features and transition-dependent states. We therefore examined the impact of the proposed system's additional feature set on the baseline system. We re-trained the proposed system without the additional energy features, and we also re-trained the baseline system as described in Section 3, but used the feature set described in Section 4.1. Results are presented in column 4 of Table 4 and column 5 of Table 5. It can be seen that not using the energy features in the proposed system resulted in a 3.8% relative increase in error at the 20-msec threshold, and that the new feature set resulted in a 3.4% relative reduction in error over the baseline system at the 20-msec threshold. These results indicate that the feature set provides a small relative improvement, but that the use of transition-dependent states and distinctive phonetic features provides the majority of the 21.24% relative improvement of the proposed work over the baseline system.

To better understand the relative contribution of each of these energy-based feature streams, we re-trained the baseline system an additional three times, with the feature set described in Section 3.1 and each of the following energy-based features: (1) burst features, (2) intensity discrimination features and their delta values, and (3) normalized energy with a 40-msec window. Adding the burst feature to the baseline system resulted in a 1.2% relative improvement over the baseline feature set at the 20-msec threshold, from 91.48% to 91.58%. Adding only the intensity-discrimination feature to the baseline feature set resulted in a 0.1% relative *decrease* in performance, from 91.48% to 91.40%. Adding only the normalized energy feature resulted in a 1.1% relative improvement at the 20-msec threshold. Therefore, the burst and normalized energy features provide small and roughly equal improvements in performance, while the intensity discrimination feature may hurt performance a small amount. We conjecture that the intensity discrimination features may not provide additional useful information in a clean recording environment, and that learning the additional parameters reduced the effectiveness of the system. The inclusion of all three features resulted in a 3.4% relative improvement over the baseline at the 20-msec threshold (from 91.48% to 91.77%), indicating that the combination of energy features contributed more than the sum of their parts.

In order to evaluate the contribution of the transition-dependent states to the proposed system performance, we tested the system without utilizing the probabilities $p(q_t \mid \mathbf{o}_t)$. In order to evaluate the contribution of distinctive-phonetic features to the system performance, we tested the system without utilizing the probabilities $p(x_t \mid \mathbf{o}_t)$. Results are given in columns 2 and 3 of Table 4. It can be seen that the system that does not utilize $p(q_t \mid \mathbf{o}_t)$ performs nearly as well as the complete system, with 93.11% agreement at 20 msec. Therefore, most of the system performance can be attributed to the transition-dependent states. The system that does not utilize $p(x_t \mid \mathbf{o}_t)$, however, performs notably worse than even the baseline system, at 88.63% agreement within 20 msec. The only difference between this system and the baseline system trained with the three energy features is the use of distinctive phonetic features instead of standard phonemic categories. Therefore, while the use of distinctive phonetic features was acceptable in the case of estimating probabilities of transition-dependent states, this aspect of the system hurt performance when estimating standard context-dependent phonetic categories. In the future, it would be interesting to combine the current transition-dependent states with standard phonetic categories.

Not all types of phoneme boundaries were improved equally using the proposed method. Comparing the proposed system with the baseline system at the 20-msec threshold, and evaluating the 20 most-frequent types of phoneme transitions, as shown in columns 3 and 5 of Table 6, there are only four cases in which the baseline system performed better than the proposed system. In three of these cases (vowel to approximant, voiced-stop to vowel, and voiced-fricative to vowel transitions), the performance was within half a percentage point. In the fourth case, for approximant to approximant transitions, the baseline system had agreement of 68.81%, and the proposed system had agreement of 67.36%. In a number of cases, the proposed system had a large relative increase in agreement. The best type of improvement was in the transitions to closure (defined as either pause or stop closure). Performance on vowel to closure transitions increased from 94.96% to 96.69%, unvoiced-fricative to closure transitions increased from 91.18% to 96.72%, nasal to closure transitions increased from 82.82% to 86.36%, approximant to closure transitions increased from 91.21% to 95.18%, and voiced-fricative to closure transitions increased from 86.90% to 94.41%. Other types of transitions also showed large relative improvement, such as closure to voiced-stop transitions (97.59% to 98.75%), and unvoiced-fricative to vowel transitions (98.72% to 99.51%), although both systems demonstrated relatively high performance. Because many of the large improvements involve phonetic classes with a large difference in energy, we also evaluated the baseline system with the additional energy features on the most frequent types of transitions. Results are given in column 4 of Table 6. While the use of energy features did help in some cases, e.g. in the transitions from vowels to closures (95.68%), in many cases the improvement due to the additional energy features was not dramatic. For example, unvoiced fricatives to closures improved only from 91.18% to 92.20%, and performance on nasal-to-closure transitions declined, from 82.82% to 80.20%. Therefore, the energy-based features did not seem to provide large benefit at transitions with especially dramatic energy changes, but they provide a smaller benefit across most types of transitions.

In order to evaluate robustness of the proposed techniques to different noise and channel conditions, we trained systems using both the baseline and the proposed techniques on the both OGI Stories corpus (Cole et al., 1995) and the TIMIT corpus. The OGI Stories corpus contains landline telephone speech of extemporaneous monologue. In order to make the data from both channels more similar, the telephone-speech data was upsampled to 16 kHz, and both sets of data were filtered with a 160-Hz high-pass filter. Each system (baseline and proposed), trained on both datasets, was then evaluated on the testing partition of each dataset. Three-fifths of the Stories corpus was used for training, and two-fifths was used for testing; training and testing partitions were speaker-independent. Results for the baseline multi-channel system are provided in columns 2 and 4 of Table 7, and results for the proposed multi-channel system are in columns 3 and 5 of Table 7. It can be seen that the proposed multi-channel system has better performance than the baseline multi-channel system on these test sets. It can also be seen that agreement of the proposed system on the TIMIT corpus (92.25% within 20 msec) is somewhat reduced compared to the system trained only on TIMIT data, and that agreement on the OGI Stories corpus (88.69% within 20 msec) is even lower than that of the TIMIT corpus. These results reflect the difficulty of processing telephone-channel speech, even though human agreement of alignments on telephone speech is about the same as on microphone speech.

7. Conclusion and Future Work

The proposed system has demonstrated a 22.1% relative reduction in error over a baseline HMM-based forced-alignment system. The result of 93.36% agreement within 20 msec is the best known reported result on the TIMIT corpus. While the choice of features accounted for some of the improvement over the baseline system, the use of transition-dependent states was responsible for the majority of the obtained agreement level.

Both the baseline system and the proposed system are available in 16-kHz and multi-channel formats to the research community, as part of the CSLU Toolkit. The CSLU Toolkit can be freely downloaded for research purposes from http://www.cslu.ogi.edu/toolkit. The systems can be run as Tcl scripts from a command line, specifying the waveform, a text file containing the list of phonemes, and the output file as command-line parameters. The scripts are in the "CSLU/Toolkit/2.0/apps/fa" directory, and are called "fa_new_16k.tcl," "fa_new_multichan.tcl," "fa_baseline_16k.tcl," and "fa_baseline_multichan.tcl," for the proposed system trained on 16-kHz data, the proposed system trained on multi-channel data, the baseline system trained on 16-kHz data, and the baseline system trained on multi-channel data, respectively. (These systems utilize the Worldbet system of phonetic symbols (Hieronymus, 1994) instead of the TIMIT phonetic symbols. For information on setting path variables in the Toolkit, see http://cslu.cse.ogi.edu/tutordemos/nnet_training/tutorial.html. Also, it should be noted that some parts of the Toolkit, e.g. the Rapid Application Developer, use older forced alignment systems and are not as accurate.) The work presented here assumes that the phoneme sequence is known, and that the only task is to determine the phoneme locations. For tasks in which only the word sequence is known, the gen_pronun.tcl script can be used to generate phoneme sequences from word sequences using a pronunciation dictionary and letter-to-phoneme rules.

Despite levels of agreement that are within a relative 2.0% of manual agreement at 20 msec, a number of research directions are still important. First, the error rate of 99.74% within 85 msec indicates a (small) number of gross errors that the system makes that a human may not make. Also, agreement levels are negatively impacted by noise and channel conditions. Speakers with dysarthria (a motor speech impairment) present special problems for a forced-alignment system, as they may have poor coordination of the articulators that results in unusual speech and timing patterns. We plan to adapt the proposed work to individual speakers, resulting in a speaker-dependent alignment system that is expected to have greater accuracy. One way in which the system performance might be improved is to model $p(x_\mathrm{t} \mid x_{t\text{-}1})$ explicitly, instead of assuming that these transition probabilities have negligible impact. Also, performance may be improved by applying the probabilities of transition-dependent states to standard context-dependent phonetic units, rather than to combinations of distinctive phonetic features. Finally, the techniques developed here can be applied to the task of automatic speech recognition instead of forced alignment.

Acknowledgments

This work was supported in part by the National Institutes of Health NIDCD grant R21-DC06722, the National Institutes of Health NIA grants AG08017 and AG024978, and the National Science Foundation grants GER-9354959 and IRI-9614217. The views expressed here do not necessarily represent the views of the NIH or NSF.

References

Bourlard H, Morgan N, Renals S. Neural Nets and Hidden Markov Models: Review and Generalizations. Speech Communication 1992;11:237–246.

Brugnara F, Falavigna D, Omologo M. Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models. Speech Communication 1993;12(4):357–370.

Campbell, N. Autolabelling Japanese ToBI. Proceedings of ICSLP '96; Philadelphia, PA. 1996. p. 2399-2402.

Cole, RA.; Noel, M.; Lander, T.; Durham, T. New Telephone Speech Corpora at CSLU. Proceedings of Eurospeech '95; Madrid, Spain. Sep. 1995 p. 821-824.

Cole, R.; Oshika, BT.; Noel, M.; Lander, T.; Fanty, M. Labeler Agreement in Phonetic Labeling of Continuous Speech. Proceedings of ICSLP '94; Yokohama, Japan. Sep. 1994 p. 2131-2134.

Cosi, P.; Falavigna, D.; Omologo, M. A Preliminary Statistical Evaluation of Manual and Automatic Segmentation Discrepancies. Proceedings of Eurospeech '91; Genova, Italty. 1991. p. 693-696.

Cosi, P.; Hosom, JP. High Performance General Purpose Phonetic Recognition for Italian. Proceedings of ICSLP 2000; Beijing, China. Oct. 2000 p. 527-530.

Cox, S.; Brady, R.; Jackson, P. Techniques for Accurate Automatic Annotation of Speech Waveforms. Proceedings of ICSLP '98; Sydney, Australia. Dec. 1998 p. 1947-1950.

Duc, DN.; Hosom, JP.; Luong, CM. HMM/ANN System for Vietnamese Continuous Digit Recognition. In: Chung, Paul WH.; Hinde, Chris; Ali, Moonis, editors. Developments in Applied Artificial Intelligence, Lecture Notes in Artificial Intelligence 2718. Berlin: Springer-Verlag; 2003. p. 481-486.

Furui S. On the Role of Spectral Transitions for Speech Perception. Journal of the Acoustical Society of America 1986;80(4):1016–1025. [PubMed: 3771921]

Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM, National Institute of Standards and Technology, NTIS Order No PB91-505065. 1990

Gillick, L.; Cox, SJ. Some Statistical Issues in the Comparison of Speech Recognition Algorithms. Proceedings of ICASSP '89; Glasgow, Scotland. May. 1989 p. 532-535.

Gong, Y.; Haton, J-P. Iterative Transformation and Alignment for Speech Labeling. Proceedings of Eurospeech '93; Berlin, Germany. 1993. p. 1759-1762.

Gordon-Salant S, Yeni-Komshian GH, Fitzgibbons PJ, Barrett J. Age-Related Differences in Identification and Discrimination of Temporal Cues in Speech Segments. Journal of the Acoustical Society of America April;2006 119(4):2455–2466. [PubMed: 16642858]

Greenberg, S. Understanding Speech Understanding: Towards a Unified Theory of Speech Perception. Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception; Keele, England. Jul. 1996 p. 1-8.

Hermansky H, Morgan N. RASTA Processing of Speech. IEEE Transactions on Speech and Audio Processing Oct;1994 2(4):578–589.

Hieronymus J. ASCII phonetic symbols for the world's languages: Worldbet. AT&T Bell Laboratories Technical Memo. 1994

Hosom, JP.; Cole, RA. Burst Detection Based on Measurements of Intensity Discrimination. Proceedings of ICSLP 2000; Beijing. Oct. 2000 p. 564-567.

Hosom JP, Cole RA, Cosi P. Improvements in Neural-Network Training and Search Techniques for Continuous Digit Recognition. Australian Journal of Intelligent Information Processing Systems Summer;1998 5(4):277–284.

Huang, X.; Acero, A.; Hon, H-W. Spoken Language Processing. Prentice-Hall PTR; Upper Saddle River, NJ: 2001.

Huopaniemi, J.; Karjalainen, M. Review of digital filter design and implementation methods for 3-D sound. Proceedings of the 102nd Audio Engineering Society (AES) Convention; Munich, Germany. March 22-25, 1997; preprint no 4461

Kain AB, Hosom JP, Niu X, van Santen JPH, Fried-Oken M, Staehely J. Improving the Intelligibility of Dysarthric Speech. Speech Communication 2007;49:743–759.

Kvale, K. On the Connection Between Manual Segmentation Conventions and 'Errors' Made by Automatic Segmentation. Proceedings of ICSLP '94; Yokohama, Japan. Sep. 1994 p. 1667-1670.

Ladefoged, P. A Course in Phonetics. Harcourt Brace College Publishers; Fort Worth, TX: 1993.

Leung, HC.; Zue, VW. A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech. Proceedings of ICASSP '84; San Diego, CA. 1984. p. 2.7.1-2.7.4.

Levinson SE. Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition. Computer Speech and Language Mar;1986 1(1):29–45.

Ljolje, A.; Hirschberg, J.; van Santen, JPH. Automatic Speech Segmentation for Concatenative Inventory Selection. In: van Santen, JPH.; Sproat, RW.; Olive, J.; Hirschberg, J., editors. Progress in Speech Synthesis. Springer-Verlag; New York: 1997.

Ljolje, A.; Riley, MD. Automatic Segmentation and Labeling of Speech. Proceedings of ICASSP '91; Toronto, Canada. May. 1991 p. 473-476.

Malfrère, F.; Deroo, O.; Dutoit, T. Phonetic Alignment: Speech Synthesis vs. Hybrid HMM/ANN. Proceedings of ICSLP '96; Sydney, Australia. Dec. 1998 p. 1571-1574.

Moore, BCJ. An Introduction to the Psychology of Hearing. Academic Press; San Diego, CA: 1997.

Pellom, BL. Ph D thesis. Duke University; Durham, North Carolina: 1998. Enhancement, Segmentation, and Synthesis of Speech with Applications to Robust Speaker Recognition.

Rabiner, L.; Juang, B. Fundamentals of Speech Recognition. Prentice Hall; Englewood Cliffs, NJ: 1993.

Rapp, S. Automatic Phonemic Transcription and Linguistic Annotation from Known Text with Hidden Markov Models: An Aligner for German. Proceedings of ELSNET Goes East and IMACS Workshop; Moscow, Russia. 1995.

Richard MD, Lippmann RP. Neural Network Classifiers Estimate Bayesian *a* posteriori Probabilities. Neural Computation Winter;1991 3(4):461–483.

Shobaki, K.; Hosom, JP.; Cole, RA. The OGI Kids' Speech Recognizers and Corpus. Proceedings of ICSLP 2000; Beijing, China. Oct. 2000 p. 258-261.

Shriberg LD, Campbell TF, Karlsson HB, Brown RL, McSweeny JL, Nadler CJ. A Diagnostic Marker for Childhood Apraxia of Speech: The Lexical Stress Ratio. Clinical Linguistics and Phonetics 2003;17(7):549–574. [PubMed: 14608799]

Shriberg LD, Green JR, Campbell TF, McSweeny JL, Scheer AR. A diagnostic marker for childhood apraxia of speech: the coefficient of variation ratio. Clinical Linguistics and Phonetics 2003;17(7): 575–595. [PubMed: 14608800]

Singh S, Bucks R, Cuerden J. Evaluation of an Objective Technique for Analysing Temporal Variables in DAT Spontaneous Speech. Aphasiology 2001;15(6):571–584.

Svendsen, T.; Kvale, K. Automatic Alignment of Phonemic Labels with Continuous Speech. Proceedings of ICSLP '90; Kobe, Japan. Nov. 1990 p. 997-1000.

Svendsen, T.; Soong, FK. On the Automatic Segmentation of Speech Signals. Proceedings of ICASSP '87; Dallas, TX. 1987. p. 77-80.

Torkkola, K. Automatic Alignment of Speech with Phonetic Transcriptions in Real Time. Proceedings of ICASSP '88; New York, NY. 1988. p. 611-614.

Wagner, M. Automatic Labelling of Continuous Speech with a Given Phonetic Transcription Using Dynamic Programming Algorithms. Proceedings of ICASSP '81; Atlanta, GA. 1981. p. 1156-1159.

Wei, W.; van Vuuren, S. Improved Neural Network Training of Inter-Word Context Units for Connected Digit Recognition. Proceedings of ICASSP '98; Seattle, WA. May. 1998 p. 497-500.

Wesenick, M-B.; Kipp, A. Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals. Proceedings of ICSLP '96; Philadelphia, PA. Oct. 1996 p. 129-132.

Wheatley, B.; Doddington, G.; Hemphill, C.; Godfrey, J.; Holliman, E.; McDaniel, J.; Fisher, D. Robust Automatic Time Alignment of Orthographic Transcriptions with Unconstrained Speech. Proceedings of ICASSP '92; San Francisco, CA. Mar. 1992 p. 533-536.

Wightman, CW.; Talkin, DT. The Aligner: Text-to-Speech Alignment Using Markov Models. In: van Santen, JPH.; Sproat, RW.; Olive, J.; Hirschberg, J., editors. Progress in Speech Synthesis. Springer-Verlag; New York: 1997.

Woodland, PC.; Leggetter, CJ.; Odell, JJ.; Valtchev, V.; Young, S. The 1994 HTK Large Vocabulary Speech Recognition System. Proceedings of ICASSP '95; Detroit, MI. May. 1995 p. 73-76.

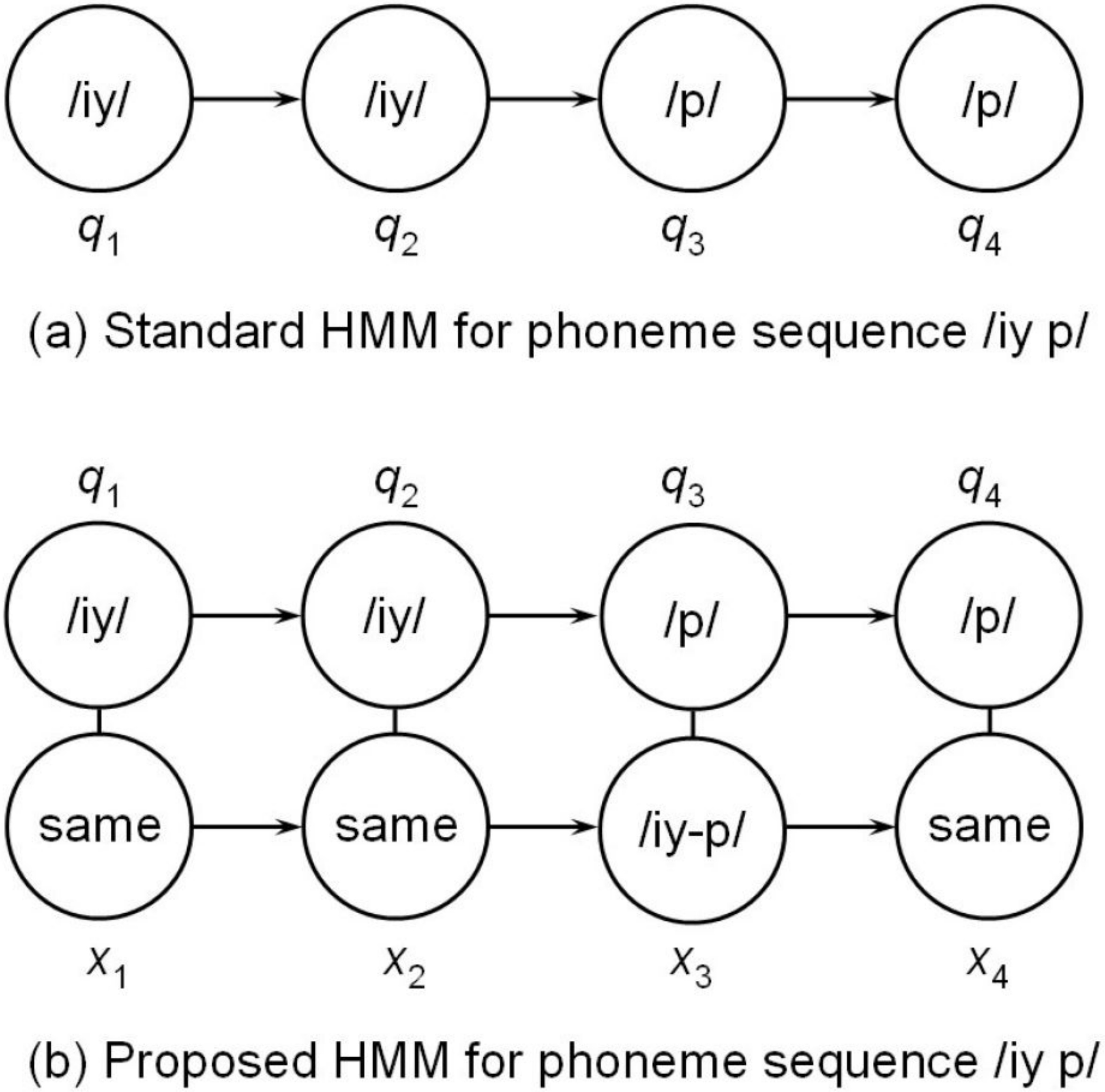

Figure 1.

Illustration of (a) a standard HMM sequence for the phonemes /iy p/ with states q_1 through q_4, and (b) the proposed HMM sequence for the same phonemes, using q_1 through q_4 and also x_1 through x_4, where the x states are linked with the q states and indicate phoneme transitions.

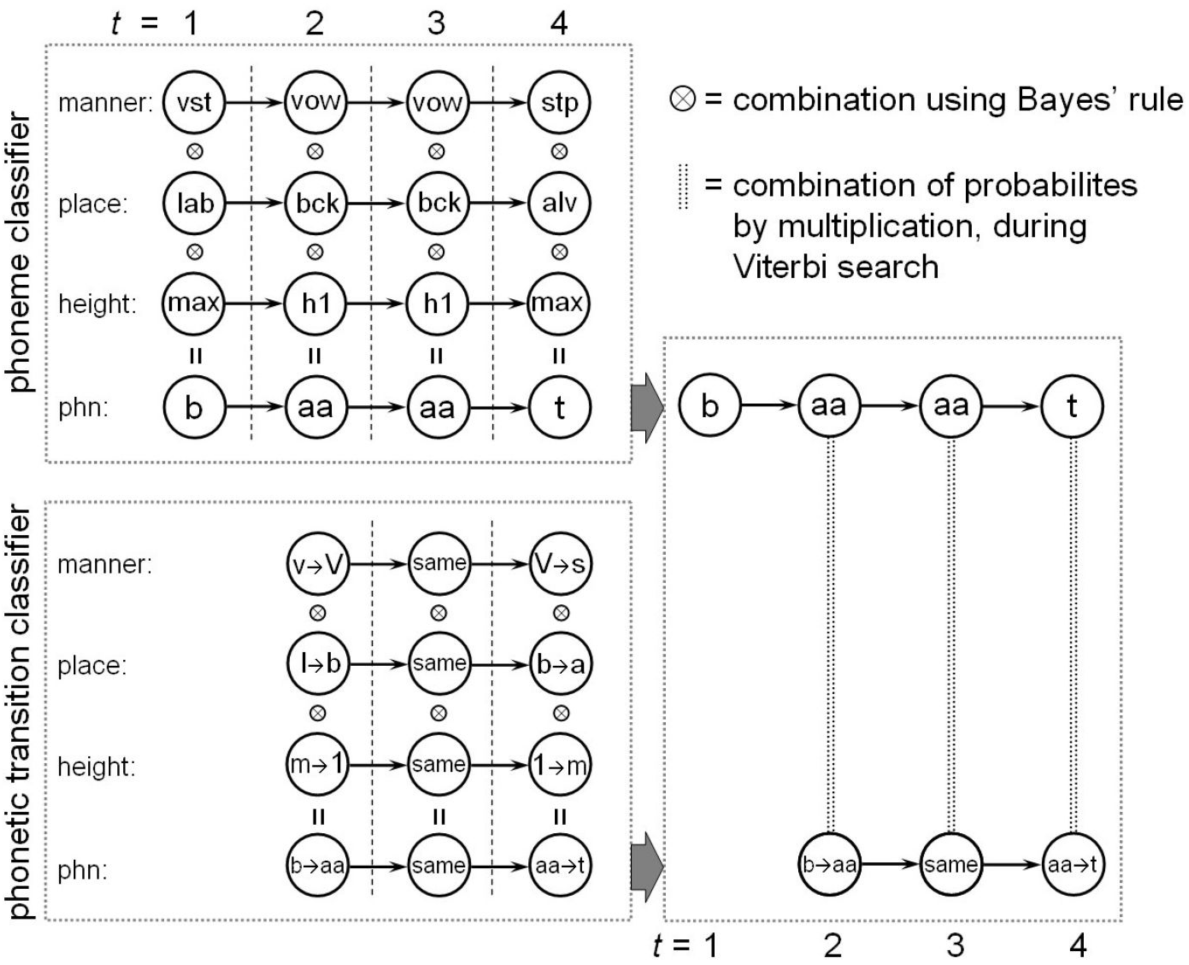

Figure 2.

Illustration of the proposed method for the phoneme sequence /b aa t/, showing how Manner, Place, and Height probabilities are combined to estimate phoneme probabilities (phoneme classifier), Manner-Transition, Place-Transition, and Height-Transition probabilities are combined to estimate phoneme-transition probabilities (phonetic transition classifier), and how these two probability sequences are combined during the Viterbi search.

Table 1

Parameters used in the baseline forced-alignment system.

Parameter Name	Value
sampling frequency	16000 Hz
frame size	5 msec
window size, type	24.0 msec, Hamming
preemphasis factor	0.97
frequency warping	Bark scale, 40 filters
cepstral weighting	0.6
feature vector	first 13 cepstral coefficients and deltas
noise reduction	low-energy cepstral mean subtraction
context window, relative to center frame	-60, -30, 0, 30, and 60 msec
phoneme set	54 phonemes
phoneme set with diphthong/affricate split	61 phonemes
context-dependent model	quasi-triphone
number of context-dependent units	451
neural network architecture	feed-forward network with 130, 300, 451 nodes in input, hidden, and output layers
number of training samples	up to 32,000 samples per category
training parameters	45 iterations, learning rate 0.05/(0.2i+1.0)
duration penalty values	0.0015 (minimum), 0.368 (maximum)

Table 2

Distinctive phonetic features (specified in the column headings) and feature values used by the proposed system.

Manner	**Place**	**Height**
vowel	front	maximum
approximant	mid	very high
nasal	back	high
aspiration	retroflex	low
unvoiced fricative	lateral	very low
voiced fricative	labial	closure
unvoiced plosive	dental	
voiced plosive	alveolar	
flap	dorsal	
closure	closure	

Table 3

Percent agreement on the test partition of the TIMIT corpus within thresholds from 5 to 100 msec, for the baseline system (column 2), the proposed system (column 3), and manual alignment (column 4). Results from manual alignment (from Section 2.2) are evaluated on a subset of 50 sentences, whereas results from the baseline and proposed system are evaluated on the entire set of 1344 test sentences. The results in columns 2 through 4 are evaluated against the canonical TIMIT phoneme boundaries. Column 5 shows results of the proposed system when evaluated against the manual alignments (50 sentences).

Threshold (msec)	**Baseline System, Percent Agreement (%)**	**Proposed System, Percent Agreement (%)**	**Manual Alignment, Percent Agreement (%)**	**Proposed System, Evaluated on Manual Alignments (%)**
5	50.78	48.28	60.38	47.96
10	76.10	79.30	81.73	79.47
15	86.45	89.49	89.07	89.46
20	**91.48**	**93.36**	**93.49**	**92.83**
25	94.27	95.38	95.36	94.76
30	96.05	96.74	96.91	95.86
35	97.25	97.61	97.79	96.80
40	98.03	98.22	98.51	97.57
45	98.58	98.62	98.79	98.45
50	98.94	98.92	99.06	98.90
55	99.19	99.13	99.50	99.06
60	99.36	99.32	99.61	99.23
65	99.47	99.45	99.67	99.50
70	99.56	99.57	99.83	99.56
75	99.62	99.64	99.83	99.61
80	99.67	99.70	99.89	99.67
85	99.74	99.75	100.0	99.78
90	99.78	99.78	100.0	99.83
95	99.82	99.81	100.0	99.94
100	99.85	99.83	100.0	99.94

Table 4

Percent agreement on the test partitions of the TIMIT corpus, within thresholds from 5 to 100 msec, for the proposed system with various modifications. In the column 2, evaluation was performed using the proposed system without $p(q_t|\mathbf{o}_t)$. In the column 3, evaluation was performed using the proposed system without $p(x_t|\mathbf{o}_t)$. In the column 4, evaluation was performed using the proposed system but without the three energy features.

Threshold (msec)	**Proposed System without p $(q_t|\mathbf{o}_t)$ (%)**	**Proposed System without p $(x_t|\mathbf{o}_t)$ (%)**	**Proposed System, No Energy Features (%)**
5	45.96	45.46	49.03
10	78.29	71.10	79.48
15	89.16	83.34	89.28
20	**93.11**	**88.63**	**93.10**
25	95.30	91.59	95.24
30	96.65	93.42	96.58
35	97.58	94.84	97.47
40	98.20	95.84	98.09
45	98.62	96.59	98.58
50	98.89	97.18	98.89
55	99.14	97.59	99.11
60	99.32	97.94	99.31
65	99.44	98.23	99.43
70	99.57	98.49	99.53
75	99.65	98.68	99.61
80	99.70	98.85	99.69
85	99.74	98.97	99.75
90	99.78	99.07	99.78
95	99.81	99.20	99.80
100	99.83	99.29	99.82

Table 5

Percent agreement on the test partitions of the TIMIT corpus, within thresholds from 5 to 100 msec, for the baseline system with the addition of selected energy features.

Threshold (msec)	**Baseline System Plus Burst Features Only (%)**	**Baseline System Plus Intensity Discrimination Features Only (%)**	**Baseline System Plus Normalized-Energy Feature Only (%)**	**Baseline System Plus 3 Proposed Features, Percent Agreement (%)**
5	51.02	49.86	49.90	51.52
10	75.75	75.57	75.59	76.11
15	86.47	86.36	86.42	86.64
20	**91.58**	**91.40**	**91.57**	**91.77**
25	94.38	94.33	94.34	94.58
30	96.07	96.16	96.09	96.25
35	97.21	97.28	97.25	97.38
40	98.03	98.07	98.04	98.08
45	98.54	98.56	98.54	98.61
50	98.96	98.96	98.95	98.98
55	99.23	99.23	99.22	99.24
60	99.44	99.40	99.38	99.42
65	99.53	99.51	99.49	99.51
70	99.63	99.61	99.59	99.62
75	99.70	99.67	99.65	99.69
80	99.75	99.73	99.72	99.74
85	99.79	99.78	99.78	99.78
90	99.84	99.81	99.82	99.82
95	99.86	99.84	99.85	99.86
100	99.88	99.86	99.86	99.88

Table 6

Comparison of baseline system, baseline system plus three energy features, and proposed system, evaluated at a 20 msec threshold on TIMIT test data, for the 20 most frequent types of phoneme transitions. The "closure" category includes both pause and stop closures.

Type of Boundary	Frequency in TIMIT Test Set (%)	Baseline System, Percent Agreement (%)	Baseline System Plus Three Energy Features, Percent Agreement (%)	Proposed System, Percent Agreement (%)
approximant – vowel	8.29	81.88	82.08	82.82
vowel – closure	8.09	94.96	95.68	96.69
closure – unvoiced stop	7.65	97.80	98.25	98.41
vowel – nasal	6.79	94.89	95.07	96.32
vowel – approximant	5.13	76.51	77.38	76.20
closure – voiced stop	4.89	97.59	98.38	98.75
unvoiced stop – vowel	4.11	97.93	98.27	98.57
vowel – unvoiced fricative	4.06	98.95	98.55	99.15
nasal – vowel	3.64	95.21	95.54	96.04
unvoiced fricative – vowel	3.33	98.72	98.90	99.51
voiced stop – vowel	3.19	99.49	99.36	99.17
vowel – voiced fricative	3.16	96.98	96.79	97.11
voiced fricative – vowel	3.06	98.07	97.94	97.81
unvoiced fricative – closure	2.79	91.18	92.20	96.72
nasal – closure	2.63	82.82	80.20	86.36
unvoiced stop – approximant	2.29	97.61	97.88	98.94
approximant – closure	2.15	91.21	90.83	95.18
vowel – vowel	1.61	70.15	72.54	74.18
voiced fricative – closure	1.27	86.90	89.78	94.41
approximant – approximant	1.26	68.81	70.90	67.36

Table 7

Percent agreement on the test partitions of the TIMIT and OGI Stories corpora within thresholds from 5 to 100 msec, for the baseline multi-channel system and the proposed multi-channel system.

Threshold (msec)	Baseline Multi-Channel System with TIMIT evaluation, Percent Agreement (%)	Proposed Multi-Channel System with TIMIT evaluation, Percent Agreement (%)	Baseline Multi-Channel System with OGI Stories evaluation, Percent Agreement (%)	Proposed Multi-Channel System with OGI Stories evaluation, Percent Agreement (%)
5	48.52	47.44	48.29	46.02
10	73.59	77.97	69.70	75.14
15	84.70	88.21	79.45	84.50
20	**90.01**	**92.25**	**84.85**	**88.69**
25	93.09	94.47	88.42	91.24
30	95.12	95.88	90.77	92.99
35	96.42	96.88	92.48	94.29
40	97.32	97.59	93.70	95.31
45	97.92	98.08	94.56	96.08
50	98.39	98.48	95.25	96.65
55	98.75	98.75	95.85	97.10
60	99.03	98.99	96.29	97.51
65	99.24	99.16	96.65	97.79
70	99.38	99.31	96.97	98.06
75	99.50	99.42	97.25	98.29
80	99.60	99.50	97.50	98.45
85	99.66	99.57	97.74	98.62
90	99.71	99.62	97.93	98.75
95	99.75	99.68	98.06	98.86
100	99.78	99.71	98.18	98.96