

Articulatory Features and Segmental Information for Automatic Speech Recognition

Torbjørn Svendsen

Department of Electronics and Telecommunications, NTNU
Trondheim, NORWAY
torbjorn@iet.ntnu.no

Abstract

The traditional speech recognizer front end possesses several weaknesses that are inherited by the subsequent processing modules. Articulatory features represent an interesting alternative to the spectrally based features normally used in automatic speech recognizers. Features describing the underlying speech production process have the potential of providing a more accurate basis for subsequent pattern recognition, and be the basis for acoustic landmark detection. Articulatory features provide an excellent basis for universal (multi-lingual) phone models, can be noise robust and can be used in combination with traditional features if so desired. Acoustic segments can be a basis for handling the asynchronous nature of the time evolution of the set of the articulatory features as well as providing an improved basis for variable frame rate analysis for obtaining accurate estimates of speech and articulatory parameters.

Index Terms: speech recognition, articulatory features, acoustic segments

1. Introduction

The performance of automatic speech recognition systems has shown good progress over the past couple of decades. However, the progress has to a large extent been driven by the availability of more data for training the statistical models and by faster and cheaper computational power, and not so much by fundamental progress in methodology, and there are indications that the performance improvement is saturating. Significant challenges need to be resolved in order to make the technology universally applicable. In general, machine word error rates are still at least one order of magnitude higher than human word error rates over a wide range of tasks[1],[2]. This implies that machine recognition of speech needs substantial improvements on all levels, from acoustic analysis and modeling, lexical modeling, language modeling and semantic analysis.

It can be argued that the main source for potential improvements are on the higher level processing. This may in particular be the case for spontaneous and conversational speech. However, the inferior machine performance also for simple recognition tasks such as digit and alpha-digit recognition, where the task performance will not benefit from syntactic or semantic information, implies that there are significant improvement potentials in the acoustic analysis and modeling. Even though humans have a remarkable capacity of decoding the spoken message even when the phonetic constituents cannot be individually identified, machines are currently far from inhibiting such capacities. Thus, the low-level acoustic processing is vital for the system performance as it will provide input to the higher level processing.

The current paper is intentionally somewhat speculative. It is the author's belief that current acoustic processing paradigms contain severe weaknesses that limit the accuracy with which speech can be automatically recognized.

2. Conventional speech recognition

The structure of a traditional speech recognizer is shown in fig. 1. In this paper we will concentrate on aspects of the front end (feature extraction).

Typically, every 10ms an analysis frame is extracted from the input speech by centering a window of 15-25ms duration at the current time instance. For each frame, speech features are extracted based on spectral analysis. The most popular speech frame representation is mel-frequency cepstral parameters (MFCC)[3], which are obtained through first estimating the log energies in a bank of triangular bandpass filters with center frequencies that are equispaced on a mel frequency scale. The vector of subband energies are then decorrelated by a cosine transform (an operation that is basically equivalent to obtaining the real cepstrum). Alternative representations, such as perceptual linear prediction (PLP)[4], which are conceptually similar, are also frequently used.

The MFCC (or PLP) representation only captures the instantaneous information in the speech signal. Important information is also contained in the temporal evolution of the spectral information. In order to capture information about the speech dynamics, approximations to the first and second order temporal derivatives of the instantaneous feature vector are normally appended to the feature vector. The output from the feature extraction is then a temporally equidistant vector sequence.

The acoustic model is typically phone oriented. Context-dependent phone models are statistical models of a phone in a given context. Each phone context is modeled by a multi-state hidden Markov model (3 states per model, and a left-right model topology is common). The emission densities of each state are typically Gaussian mixture models, GMMs. In order to reduce the number of free parameters, state tying is used, i.e. sharing model parameters between states of the context dependent phone models. The link between the phone level description and the word level is given by the pronunciation lexicon, which typically provides a canonic, broad phonetic description of the pronunciation of each vocabulary word. It is also possible to have pronunciation lexica which include pronunciation variants, often with some weight reflecting the frequency of occurrence of the variant.

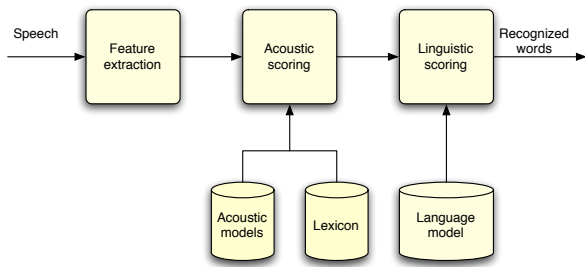


Figure 1: *Conventional speech recognizer*

3. Which features?

The purpose of the ASR front end is to extract all necessary information for the task of discriminating sounds, words and utterances in a manner that is maximally robust to variations irrelevant to the speech recognition, e.g. ambient and channel noise, speaker characteristics. The solution in current speech recognizers invariably rely on short-time spectral analysis (including temporal derivatives and often some (normalized) energy measure). All subsequent processing is reliant on the quality of the front end analysis. It is our belief that the current approach has inherent weaknesses, and that it is vital for further improvements in speech recognition that the ASR front end is carefully re-examined.

The standard front end approach has several weaknesses:

1. The spectral analysis is performed at temporally equidistant instances, using a fixed length window, and is based on the assumption that within the analysis window, the signal can be viewed as quasi-stationary. The time shift and window length are chosen as a compromise between the need for sufficient data samples to obtain reliable estimates and the need for temporal resolution to capture the acoustic variability in the speech signal. Of course, the stationarity assumption will at times be erroneous, e.g. when the analysis window is placed at phone transitions or inside phones that are not acoustically stable, such as plosives and diphthongs.
2. The temporally equidistant analysis produces a sequence of feature vectors that in HMM-based ASR are assumed to be conditionally independent. This assumption is well known to be flawed.
3. The same analysis is applied regardless of the properties of segment under consideration. It is conceivable that better analyses could be obtained if were possible to perform signal dependent analysis to the input speech signal.

3.1. Articulatory features

An alternative to the traditional, spectrally based, features which aim at describing the realized speech signal, is features that are designed to describe the underlying speech production process. Such features can emanate from an articulatory model, a model describing the interaction between the articulators involved in speech production. A popular model is based on Chomsky and Halle's distinctive feature theory[5]. The distinctive features are a set of binary valued attributes describing aspects of articulation and perception. As such, the articulatory

features have a set of properties making the attractive for speech recognition such as

- The potential of being largely language independent[6].
- Robustness to noise[7]
- The possibility of being combined with traditional features.

The sets of articulatory features that have been used for speech recognition vary. But in order to be appropriate for ASR, they must have the following properties[8]:

- correlate well with the acoustic signal to enable machine learning from data
- provide a compact representation
- encode the distinctions necessary for word discrimination

In addition, it is necessary that the features can be reliably detected from the acoustic signal.

An example of an articulatory feature set[9] is shown in Table 1.

Attribute	Allowed values
Sonority	vowel, obstruent, sonorant, syllabic, silence
Voice	voiced, unvoiced, n/a
Manner	fricative, stop, flap, nasal, approximant, nasalf flap, n/a
Place	labial, dental, alveolar, palatal, velar, glottal, lateral, rhotic, n/a
Height	high, mid, low, lowhigh, midhigh, n/a
Front	front, back, central, backfront, n/a
Round	round, nonround, roundnonround, noroundround, n/a
Tense	tense, lax, n/a

Table 1: *Articulatory features: Attributes and their values (from [9])*

In practice, detection of the articulatory features cannot be perfectly undertaken. Depending on the type of feature, the frame-level detection error rates are typically in the 10-25% range[8],[10]. Because of this, making hard decisions on the frame level (forcing detector outputs to 1 or 0) is clearly infeasible. Estimators of attribute posteriors will facilitate using the articulatory features in a subsequent statistical decoding framework. Employing a redundant set of features with respect to enabling phone and word identification can increase robustness and aid to achieve the desired performance goals.

It is worth noting that a phonemic transcription is not sufficient for providing the ground truth of the articulatory features. Allophonic variations, e.g. due to the phonetic (and articulatory) context will influence whether a canonic articulatory feature is realized or not. An example is the phone /h/ which is generally unvoiced, but in some contexts are realized as voiced. Because of this, it is necessary that articulatory feature detectors are trained using transcriptions that take into account the phonetic and articulatory context. If transcriptions are not available, and the phone sequence must be generated from pronunciation lexica (which generally contain only a broad phonetic transcription) a finer detail must be predicted e.g. by incorporating phonological rules. In traditional phone-based HMM recognizers the context dependency of the acoustic realizations are of course handled by training context dependent phone models.

Another aspect that needs to be taken into consideration is that the articulatory features will not remain constant over an entire phone. Also, the features may change asynchronously, such that a change in e.g. the voicing attribute will not necessarily occur simultaneously with the change in other features. Clearly, the ordinary HMM framework will not be well suited to handle such phenomena. In [8] it is proposed to employ dynamic Bayesian networks as a mechanism to model the asynchrony between features.

A framework for attribute-oriented speech recognition is depicted in Figure 2. The speech signal is input to a bank of attribute detectors, each providing a temporal stream of posterior probabilities for the occurrence of a given attribute. The output from the attribute detectors are combined in the event merger, to produce phone or phone state posteriors. Finally, the estimated evidence is employed to decode the phone string in the evidence verifier. A system based on this architecture, using 22 articulatory features, has been employed for continuous phone recognition on the TIMIT database yielding a phone accuracy rate of 75.0%, outperforming HMM recognizers and being competitive with the best results reported[11].

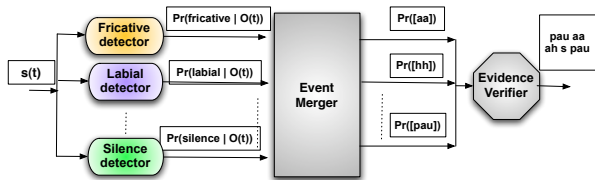


Figure 2: Attribute driven speech recognizer

3.2. Acoustic segmentation

We propose to introduce a pre-analysis step which performs an identification of segments in a speech utterance with acoustically well defined properties. A segmentation into varying duration segments can be able to provide:

1. A means for determining an optimal trade-off between temporal resolution and the number of data points available for estimation if the segmentation aims at identifying stationary segments, as the statistical properties of the segment will be near constant.
2. Reduced statistical dependency between successive feature vectors if each segment is represented by a single feature vector.
3. A possibility of applying signal dependent analyses, e.g. dependent on segment duration, or global properties of the segment.
4. A means for analysis which takes into account the asynchronous behaviour of the articulatory features and can optimize the detection of each individual feature.

Recently, there has been a renewed interest in event-based speech recognition[12]. Speech events are a collection of fundamental speech properties such as acoustical landmarks and distinctive features, and can be combined to detect phonemes, words and sentences. The distinctive features constitute a set of acoustic-phonetic features that, if properly detected, can discriminate between speech sounds. However, detection of the distinctive features by using traditional, frame based analysis,

can be unreliable, partly due to the variability in duration of the events.

The mel-frequency cepstral coefficients produced by current front ends have proven to be fairly robust and efficient. Regardless of their success, the analysis represents an ignorance paradigm, as no speech specific information is utilized in the analysis (the mel scale is a feature of hearing, not of speech, although it might be a matter of discussion to what degree the evolution of speech production and hearing have had mutual influence). On the other hand, a large body of speech feature analysis methods has been used in speech science and speech analysis. Relevant examples include estimation of voicing, formants, glottal pulses, nasality and voice onset time (VOT). Such features could be particularly relevant for estimation of distinctive features, and would potentially benefit from an initial segmentation of the acoustic signal. Such segmentation can be performed for each individual feature, producing asynchronous streams of segmental information, or be required to synchronize across all, or subsets of features.

The speech community has proposed a number of approaches to speech segmentation, see e.g. [13],[14], [15], [16], [17], [18]. Most of the research has been aimed at producing segmentations that align with phones. In the current context, we are more concerned with a segmentation that will facilitate the detection of articulatory events and acoustic landmarks. Several methods for performing acoustic segmentation have been proposed, see e.g. [17], [18]. We will present some examples based on an extended version of Constrained Clustering Segmentation[16]. The basic segmentation algorithm is presented below:

Assume that a recorded utterance is represented by a sequence of feature vectors $\{\mathbf{x}(0), \mathbf{x}(1), \dots, \mathbf{x}(T)\}$. Generally, the task of CCS can be formulated as finding the set of segment boundaries $\{b_0 = 0, b_1, b_2, \dots, b_{J-1}, b_J = T\}$ that minimizes the total distortion

$$D_{tot} = \sum_{j=0}^{J-1} \sum_{n=b_j}^{b_{j+1}-1} d(\mathbf{x}(n), \hat{\mathbf{x}}_j(n - b_j)) \quad (1)$$

where $\hat{\mathbf{x}}_j(k)$ is the approximation of the speech frames of the j 'th segment. The approximation can in principle be chosen freely, e.g. linear evolution over the segment ($\hat{\mathbf{x}}_j(k) = k * \Delta \hat{\mathbf{x}}_j + \hat{\mathbf{x}}_j$), parabolic evolution ($\hat{\mathbf{x}}_j(k) = k^2 * \Delta^2 \hat{\mathbf{x}}_j + k * \Delta \hat{\mathbf{x}}_j + \hat{\mathbf{x}}_j$) and so on. For the purpose of identifying acoustically stationary segments we choose an approximation that is constant over the segment:

$$\hat{\mathbf{x}}_j(k) = \hat{\mathbf{x}}_j \quad (2)$$

Similarly, the distortion measure, $d(*, *)$ will need to be chosen to give a reasonable interpretation. For many speech features, the Euclidian distance is a simple and meaningful choice. E.g., when using cepstral features, the Euclidian distance in the cepstral space is a good approximation to the L_2 spectral distance:

$$d(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|^2, \quad (3)$$

In this case, the approximation $\hat{\mathbf{x}}_j$ of segment j which minimizes the segment distortion is simply the mean of the feature vectors contained in the segment.

The acoustic segmentation will in principle require a minimization of eq. (1) over all possible boundary combinations. However, the algorithm can be efficiently implemented by utilizing the dynamic programming level building principle[16].

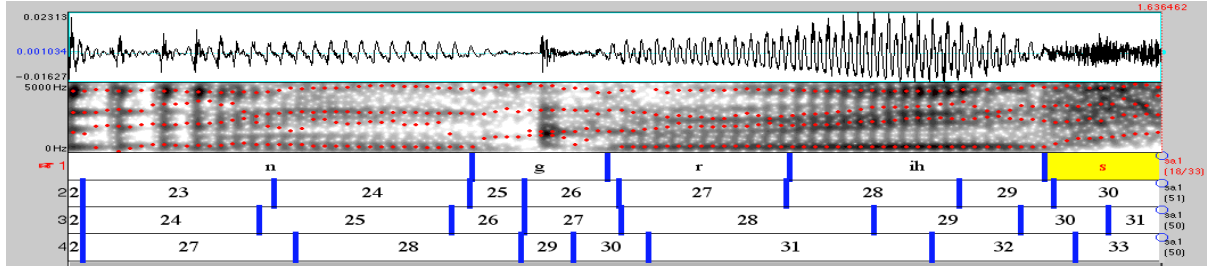


Figure 3: Acoustic segmentation and the corresponding manual phonemic annotation of a section of the sentence "She had your dark suit in greasy washwater all year". From bottom is 2nd, 1st and 0th order segmentation.

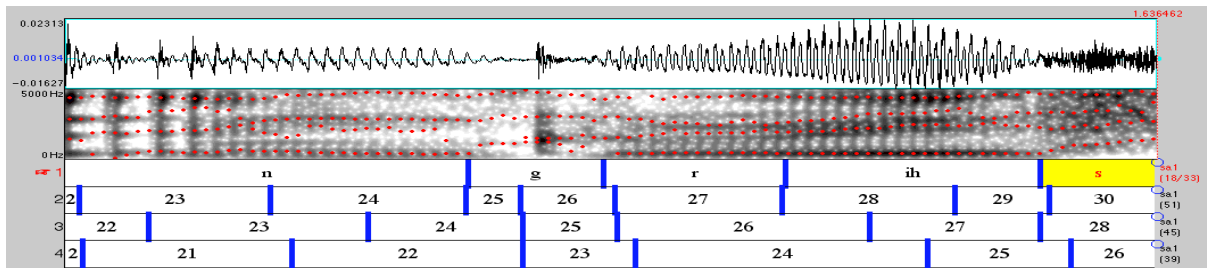


Figure 4: Acoustic segmentation and the corresponding manual phonemic annotation when the number of 1st and 2nd order segments are 10% more than and equal to the number of phones in the sentence, respectively.

Noting that the initial boundary b_0 will be the start of the utterance and b_J corresponds to the utterance end, we start by finding the internal boundary that optimizes eq. (1) for two segments. Results from the 2-segment solution can then be employed in the search for the best 3-segment solution and so on. The number of acoustic segments, J , for an input speech utterance is unknown. We have used a threshold on the average distortion as the criterion for selecting J , i.e. J is the smallest number of segments such that $D_{tot}(J)/b_J \leq \Theta$. The choice of Θ will depend on the feature vector and the distortion measure employed.

An example of acoustic segmentation of a fragment of a TIMIT sentence is shown in Figure 3. The uppermost annotation layer shows the manual segmentation. The three annotation layers below show (top to bottom) segmentations for the zeroth order, first order and second order approximations. In this example, the number of segments for each approximation is kept approximately equal, producing $\sim 38\%$ more acoustic segments than phones on the sentence level. It can be seen that the 0th and 1st order segmentation segments the closure and burst phase of the /g/ in separate segments, while the parabolic segmentation includes the closure into the last part of the /n/. The gradual spectral change in the section containing the phones /r/ and /iy/ is also handled differently. The constant approximation returns segments that correspond well with the manual segmentation. The 1st order segmentation finds a segment containing the /r/ and the initial part of the /iy/ which can be well modeled by a linear approximation.

In Figure 4 the number of segments for the first and second order approximation is reduced to accommodate the greater modeling capacity of the higher order approximations. The linear approximation has 20% more segments and the parabolic segmentation has only 5% more segments than phones in the sentence. In this case we see that the differences between the resulting segmentations are even more pronounced.

These examples illustrate that the choice of segmentation approach will yield segments with different properties. The piecewise constant approximation will aim to return segments where the observations are stationary. Transitional areas will typically be represented as short segments, since this will minimize the approximation error. The piecewise linear approximation will favor segments with constant or slowly varying properties. All approximations will tend to place segment boundaries at places of abrupt change.

3.3. Segment oriented processing

The acoustic segmentation can be employed as a purely pre-processing step, in order to improve the performance of the articulatory feature detectors. Each feature detector can have feature dependent input observations. E.g. in [19], features like zero crossing rate, probability of voicing, subband energy maxima and spectral dynamics were used as acoustic correlates for various articulatory features. In calculating the input observations from the acoustic waveform, the segmentation information can be utilized to improve the reliability of the estimates.

The articulatory features and an associated segmentation can furthermore be combined to detect acoustic landmarks[20] which can provide vital cues for automatic speech recognition, and can be used to synchronize multiple classifiers.

Finally, the use of segments provide a link to segment models (see e.g. [21], [22]), although an explicit segmentation is not usually required in a segment model. Actually, in [23], a direct link between the segmentation algorithm used to exemplify acoustic segmentation (zeroth order segmentation) and constant trajectory segment models was identified.

4. Conclusions

Articulatory features represent an interesting alternative to the spectrally based features normally used in automatic speech recognizers. Features describing the underlying speech production process have the potential of providing a more accurate basis for subsequent pattern recognition, and be the basis for acoustic landmark detection. In addition, studies have shown that articulatory features provide an excellent basis for universal (multi-lingual) phone models[6], and can be noise robust. Acoustic segments can be a basis for handling the asynchronous nature of the time evolution of the set of the articulatory features as well as providing an improved basis for variable frame rate analysis for obtaining accurate estimates of speech and articulatory parameters.

5. Acknowledgements

This work has been supported by the Research Council of Norway's VERDIKT programme through the SIRKUS project.

6. References

- [1] R. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, pp. 1–15, 1997.
- [2] D. Pallett, "A look at nist's banchmark asr tests: past, present and future," in *Proc. IEEE ASRU Workshop 2003*. IEEE, 2003, pp. 483–488.
- [3] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP 28, no. 4, pp. 357–366, 1980.
- [4] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.
- [5] N. Chomsky and M. Halle, *The Sound Pattern of English*. Harper&Row, 1968.
- [6] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Toward a detector-based universal phone recognizer," submitted to ICASSP'08.
- [7] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *Proc. ICSLP'98*, Sydney, 1998, pp. 891–894.
- [8] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic bayesian networks," *Computer Speech and Language*, vol. 21, pp. 620–640, 2007.
- [9] J. Morris and E. Fosler-Lussier, "Discriminative phonetic recognition with conditional random fields," in *HLT-NAACL Workshop on Computationally Hard Problems and Joint Interference in Speech and Language Processing*, 2006.
- [10] K. Hacioglu, B. Pellom, and W. Ward, "Parsing speech into articulatory events," in *Proc. ICASSP'04*, 2004, pp. I-925 – I-928.
- [11] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Towards bottom-up continuous phone recognition," in *Proc. ASRU'07*, Kyoto, Dec. 2007.
- [12] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: a new speech research paradigm for next generation automatic speech recognition," in *Proc. ICSLP'04*, 2004.
- [13] R. Obrecht, "A new statistical approach for the automatic segmentation of continuous speech," *IEEE Trans. Acoust. Speech Signal Processing*, Jan 1988.
- [14] A. Sethy and S. Narayanan, "Refined speech segmentation for concatenative speech synthesis," in *Proc. ICSLP'02*, Denver, Sept. 2002, pp. 149–152.
- [15] J. Glass and V. Zue, "Multi-level acoustic segmentation of continuous speech," in *Proc. ICASSP'88*, 1988, pp. 429–432.
- [16] T. Svendsen and F. K. Soong, "On the automatic segmentation of speech," in *Proc. ICASSP'87*, 1987.
- [17] A. Brandt, "Detecting and estimating parameter jumps using ladder algorithms and likelihood ratio tests," in *Proc. ICASSP'83*, 1983.
- [18] V. Tyagi, H. Bourlard, and C. Wellekens, "On variable-scale piecewise stationary spectral analysis of speech signals for asr," *Speech Communication*, vol. 48, no. 9, pp. 1182–1191, Sept. 2006.
- [19] N. N. Bitar and C. Y. Espy-Wilson, "Knowledge-based parameters for hmm speech recognition," in *Proc. ICASSP'96*, 1996, pp. 29–32.
- [20] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan, J. Muller, K. Sonmez, and T. Wang, "Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop," in *Proc. ICASSP'05*, vol. 1, 2005, pp. 213–216.
- [21] M. Ostendorf, V. Digalakis, and A. Kimball, "From hmm's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech and Audio Proc.*, vol. 4, no. 5, pp. 360–378, 1997.
- [22] M. Russel and W. Holmes, "Linear trajectory segmental hmm's," *IEEE Signal Processing Letters*, vol. 4, no. 3, pp. 72–74, 1997.
- [23] M. Russell, "A segmental hmm for speech pattern processing," in *Proc. ICASSP'93*, 1993, pp. II-499–II-592.