On speech variation and word type differentiation by articulatory feature representations

Louis ten Bosch, Harald Baayen, Mirjam Ernestus

CSLT
Radboud University, Max Planck Institute for Psycholinguistics, Nijmegen, NL
l.tenbosch@let.ru.nl, harald.baayen@mpi.nl, mirjam.ernestus@mpi.nl

Abstract

This paper describes ongoing research aiming at the description of variation in speech as represented by asynchronous articulatory features. We will first illustrate how distances in the articulatory feature space can be used for event detection along speech trajectories in this space. The temporal structure imposed by the cosine distance in articulatory feature space coincides to a large extent with the manual segmentation on phone level. The analysis also indicates that the articulatory feature representation provides better such alignments than the MFCC representation does. Secondly, we will present first results that indicate that articulatory features can be used to probe for acoustic differences in the onsets of Dutch singulars and plurals.

Index Terms: articulatory features, automatic speech recognition, speech decoding, lexical differentiation.

1. Introduction

Virtually all approaches in automatic speech recognition (ASR) systems assume that the information in the speech signal and ASR dictionaries can be represented in terms of sequences of discrete symbols (e.g. phone-like symbols). This beads-on-a-string paradigm ([9], which goes back to e.g. [16]), forces a less than optimal representation of variation in speech since variation (due to pronunciation variation, speaking styles, inter-speaker differences, accents etc.) primarily takes place in a continuous domain, often with effects on the sub-phonemic level, rather than in a discrete domain. The description of variations in a continuous domain by discrete symbols is evidently a result of compromises (cf. [1]). It can therefore be argued that fundamentally better ways to model variation in speech can be achieved by modeling the underlying pronunciation process rather than modeling the surface effects on the resulting acoustic speech signal. In this area, progress has been made by using specifically trained articulatory feature classifiers ([4], [6], [10], [12], [7]). The choice of the set of articulatory features is largely inspired by both the theory of distinctive features ([3]) and the gestural theory of speech production ([2]).

In this study, we describe ongoing research that aims at a description of the variation in speech by use of articulatory features (AF). As in [4], [18], we apply AF classifiers using a feature set including manner of articulation, place of articulation, voicing, front-back and rounding. The combination of the AFs results in a sequence of vectors (updated each 10 ms) defining a trajectory in AF space.

Compared to other representations, AFs offer two advantages. First, AFs provide a description of the speech signal allowing loose synchrony between articulatory features, in contrast with linear phone representations which explicitly impose strictly synchronous feature transitions. Secondly, AFs make it possible to provide a strong link between variation in speech and the relevance of fine phonetic details in human speech processing. There is a growing number of indications that human lexical decoding is mediated by subphonemic details (e.g. [15], [19]).

In this paper, we will first describe a framework in which AFs are used for event detection. We will use the term event to mark a salient, sudden change in the trajectory. Trajectories are endowed with a temporal structure by using a distance in AF space, and we will compare this structure with manual segmentation. Another approach to impose a structure on the bottom-up derived AF streams is based on Dynamic Bayesian Networks DBN (e.g. [13]) or on parsing [5]). The method presented here can be regarded as an alternative and complementary way to relate event-detection and structure in the speech signal.

Secondly, we will describe experiments showing how articulatory features can be used to distinguish different word types. Recently, studies have observed systematic differences in acoustic duration between words in isolation (e.g., *ham*) and the same words embedded in longer words (e.g., *hamster*) [19, 23]. In Dutch, the duration of a syllable is dependent on the number of syllables that follow in the word and may therefore mediate word differentiation. [21, 22] have shown that such durational differences indeed bias the listener's interpretation. Already before the vowel of the suffix is actually perceived, listeners perceive whether a singular or a plural is involved. Our modeling experiments show that AFs form a powerful and interpretable representation in the computational modeling of similar effects.

The organisation of this paper is as follows. The next section is devoted to a brief introduction to the design and training of the AF classifiers. The third section describes the database of spontaneous speech that was used in this study, while the fourth section discusses two applicatons: event detection by distance measures in AF space, and the use of AFs in the modeling of perceptual differences on the basis of word onsets. The final section concludes with a discussion and remarks for further research.

2. Articulatory Feature Classifiers

In line with current approaches in this area (e.g. [4]), articulatory features are derived from the signal by using Artificial Neural Nets (ANN). For the ANNs used in this paper, we applied the NICO-toolkit ([11]). Each of the six features (manner, place, front-back, voicing, rounding, and static (see table 1) is represented by one ANN. Each ANN is trained on canonical feature transcriptions on the basis of a phoneme transcription of the speech signal and a phone-to-feature table. Put in paral-

Table 1: The six features with the 28 values used in this study.

Features	Card	Values
manner	6	approximant, fricative, nasal, stop, vowel, silence
place	7	(labio)dental, alveolar, velar, high, mid, low, silence
voicing	3	voiced, voiceless, silence
rounding	4	rounded, unrounded, nil, silence
front-back	5	front, central, back, nil, silence
static	3	static, dynamic, silence

lel, the six AF classifiers provide information without any imposed structure: the strict dependency of the features observed on the canonical training samples is lost due to independency between the classifiers, and so on a test set AF output vectors may show feature asynchrony and deviate from the canonical 0/1 AF vectors. The AF output consists of 28 parallel analog values between 0 and 1, updated every 10 ms.

3. Database description

In this study, we have used the IFAcorpus ([14]), a database of spoken Dutch. It contains recordings of 4 male and 4 female speakers, varying from 15 to 66 years in age. For all utterances, manually corrected labelling and segmentation on phone and word level are available. Metadata include education level, birth place, and smoking habit and contain more information than is available in the much larger Spoken Dutch Corpus (CGN, [8]). The transliteration of the IFAcorpus is according to the CGN-protocol. Compared to CGN, the amount of speech per speaker is much larger (40 min/speaker) and more speaking styles have been recorded (8, varying from spontaneous monologues to read-aloud word lists). A number of 19867 utterances have been transcribed (a bit more than 5 hours). Two subcorpora (retold stories in the form of long monologues, and randomly presented sentences, in total about 140 minutes of speech) have been selected for this study. The total number of utterances is 2650. All speech material has been converted to 16 kHz 16 bits/sample mono wav files. The phone alphabet was cleaned up to contain 50 different phones apart from the basic phones, the IFAcorpus also uses palatalised variants. There is one silence symbol. The selected subcorpus was divided into a training set (1978 utterances), a validation set (100) and a test set (572 utt; 44m10s). The test set consisted of the speech from one male and one female who were kept separate, while speech from the other 6 speakers was used for training and validation.

The training and validation set have been applied for the training of the six different ANNs. Table 2, second column, shows the classification results on the IFAcorpus test set (the accuracy of the individual classifiers on frame level in percentage correct). For the sake of comparison, we added the ANN results obtained on the TIMIT test set after training on the TIMIT training set, but since transcription methods and database specifications differ in detail, a further cross-database comparison hardly makes sense. After training, the classifiers were used to produce AF vector sequences for test data, overall resulting in about 265000 vectors of dimension 28.

Table 2: Frame-based accuracy of individual feature classifiers (in perc.) on the IFA-corpus and TIMIT test set.

Features	IFA-corpus	TIMIT
manner	84.7	86.5
place	76.7	78.6
voicing	93.5	92.0
rounding	87.4	86.0
front-back	83.6	83.0
static	89.7	81.0

Table 3: Alignment results for three distances and two signal representations. Corresponding thresholds are indicated between brackets. For an explanation see the text.

repr.	cosine	Euclidean	city-block
AF	40, 89, 3.1 (0.21)	39, 88, 3.2 (0.20)	34, 83, 5.1 (0.41)
MFCC	34, 79, 6.1 (0.85)	32, 81, 9.0 (325)	35, 79, 4.9 (106)

4. Two applications of AF representations

4.1. Bottom-up alignment with manual segmentations

Given a certain distance function D, event along a trajectory $\ldots, v_{n-1}, v_n, v_{n+1}, \ldots$ may be defined by the moments on which $D(v_{i-1}, v_i)$ exceeds a certain threshold θ (of which the optimal value depends on the type of distance). It has been shown ([18]) that this technique yields promising alignment results between events and manual phone-level segmentations when D equals the cosine distance (eq. 1) and the v_i represent AF vectors.

$$D = \arccos \frac{(v^{(1)}, v^{(2)})}{|v^{(1)}||v^{(2)}|} \quad (1)$$

This difference is further elaborated in table 3. The table shows alignment results between the event detection and the manual segmentation for three distances (cosine, Euclidean, city-block) and two representations (MFCC, AF). The figures indicate the percentage of frames with an exact match, with an match within 25 ms, and without a manual segment boundary within 5 frames, respectively. The optimal threshold θ is given between brackets. For example, for the combination (cosine, AF), 40 percent of the cosine peaks coincide with the manual boundary, while 89 percent could be assigned a boundary within 25 ms from the cosine peak, and 3.1 % (215 out of 6810 cosine-maxima) could not be associated with a segment boundary within the range [-5, 5]. For all other combinations the alignment is worse, but the Euclidean distance performs almost equally well. The value of 89 percent within 25 ms is comparable to the accuracy of 84 percent within 20 ms (reported in [17], table 5) for the position of phone boundaries by automatic segmentation.

This alignment result is not coincidental. Theoretically, it *might* be the case that all point processes with a similar statistics as the manual segmentation can be aligned with the same success rate. This possibility, however, turns out to be less likely since (a) it appears that the manual segmentation distribution is very similar to a Poisson distribution with $\lambda = 5.7$ (see figure 1),

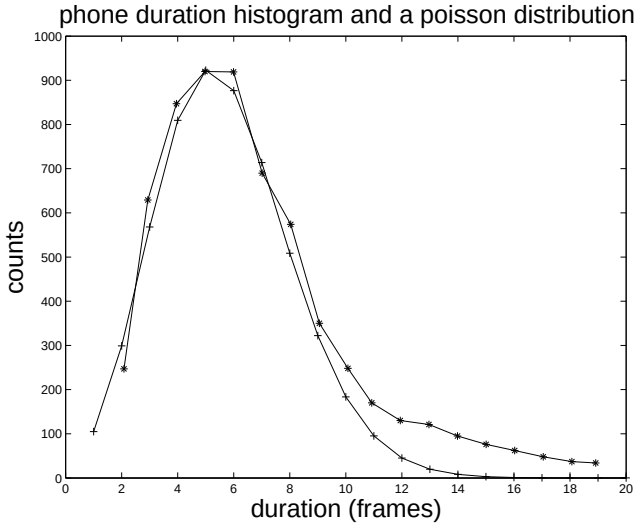

Figure 1: Histograms of segment durations (stars) and the Poisson distribution ($\lambda = 5.7$) (+-signs).

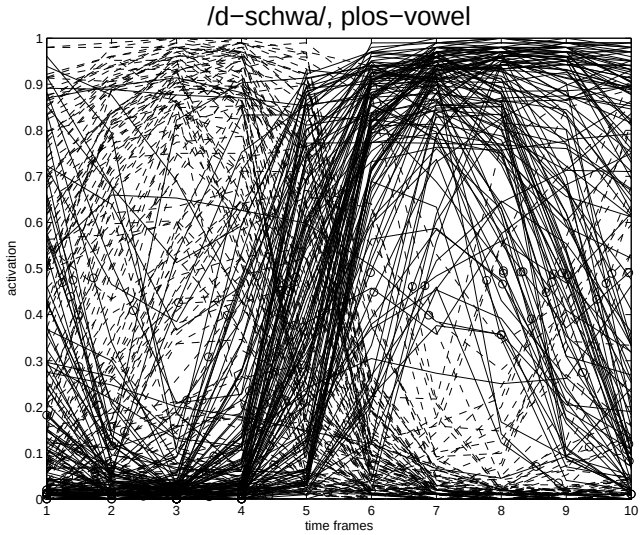

Figure 2: Feature values (plosive, vowel) over time for /d/-schwa.

and (b) the alignment between the cosine peaks and a Poisson process with $\lambda = 5.7$ is significantly different from the alignment between the cosine peaks and the manual segmentation ($\chi^2 = 11.6$, nf $= 2$, $p < 0.01$).

In order to apply AF representation to study fine temporal-phonetic details, probably a time resolution finer than 10 ms is required. This is suggested by figure 2, which shows over 80 realisations of the transition d-schwa. Increasing (solid line) and decreasing (dashed) plots display the feature value 'vowel' and 'plosive', respectively. All plots are overlaid such that the *manual* segment boundary is halfway between frame 5 and 6. The coarse resolution due to the 10-ms time frame shift is clearly visible.

4.2. Relevance of fine phonetic details in word onsets

As mentioned above, listeners can perceive subtle differences between words in isolation (e.g., *ham*) and the same words embedded in longer words (e.g., *hamster*) [19, 23]. In order to investigate how differences other than duration differentiate Dutch singulars and plurals, an experiment was conducted using AF representations of acoustic realisations of Dutch singulars and plurals.

4.2.1. Materials

For 47 nouns, we recorded several tokens of the singular and plural form, read by a female native speaker of Dutch. All plurals were bisyllabic words ending in the plural suffix *-en*, all singulars were monosyllabic. The number of singular tokens ranged from 2 to 5, the number of plural tokens ranged form 14 to 20. In all, 993 tokens were recorded, and digitized at a sample rate of 44 kHz. For each token, the corresponding matrix of scores was calculated. The average duration of a token was 54 timesteps. The plural forms (range 35-85) tended to be shorter than the singular forms (range 23-97) by 5 timesteps (i.e., by approximately 50 ms, $p < 0.0001$, mixed-effect anova with word as random stratum). For each word, 2 tokens of the singular and 14 tokens of the plural form were randomly selected for training, the remaining tokens (maximally 3 singulars, 6 plurals) were held out for testing.

4.2.2. Results

We fitted a stepwise logistic regression model to the data with the log odds of plural to singular as the dependent variable, and the acoustic feature values as predictors. An initial mixed-effect analysis with Word as random effect revealed negligeable variation for this factor ($\hat{\sigma} < 0.0001$). We therefore removed Word as a predictor from the model. A stepwise logistic regression analysis suggested significant predictivity for the feature values `labiodent` ($\hat{\beta} = 11.9$, $\hat{\sigma} = 3.18$, $Z = 3.73$, $p = 0.0002$), `unvoiced` ($\hat{\beta} = 2.570$, $\hat{\sigma} = 1.0413$, $Z = 2.47$, $p = 0.0136$), `voiced` ($\hat{\beta} = 0.782$, $\hat{\sigma} = 0.3007$, $Z = 2.60$, $p = 0.0093$), `back` ($\hat{\beta} = 2.336$, $\hat{\sigma} = 0.8403$, $Z = 2.78$, $p = 0.0054$), `static-nil` ($\hat{\beta} = 3.156$, $\hat{\sigma} = 0.4662$, $Z = 6.77$, $p = 0.0000$), and `static` ($\hat{\beta} = 1.701$, $\hat{\sigma} = 0.4282$, $Z = 3.97$, $p = 0.0001$). Although significant, the features succeeded in accounting for only a tiny proportion of the variance. The bootstrap-corrected R^2 was 0.025 and the bootstrap-corrected value of Somers $D_{xy} = 0.21$. All 6 predictors were retained in 156 out of 200 bootstrap runs using a backwards variable elimitation algorithm [20].

A t-test on the predicted probabilities for the singular and plural revealed a highly significant $p < 0.0001$ difference in probability of 2% (mean predicted probability singular: 0.86, mean predicted probability plural: 0.88). When applied to the held-out singulars and plurals, a t-test on the predicted probabilities for the held-out singulars and plurals revealed a significant ($p < 0.0001$) difference in probability of 1% in the expected direction (mean predicted probability singular: 0.87, mean predicted probability plural: 0.88).

These results suggest that there are subtle qualitative differences in the fine phonetic detail in the first 50 ms of Dutch singulars and plurals. The plurals in our data appear to have been realized with more acoustic detail for labio-dental place of articulation, more detail for voicing, and more to the back of the mouth. The evidence for staticity is mixed, with one feature value indicating that statisticity is irrelevant for plurals (`Snvt`)

and another feature value indicating more evidence for staticity (`Sstatic`). Considered jointly, the main pattern is that plurals receive more careful articulation than singulars. In the light of the shorter duration of the stem in plurals, this suggests that durational shortening is compensated for by increased articulatory detail.

5. Discussion and conclusion

We addressed variation in speech by aligning data-based events with manual phone segmentations and by relating acoustic details in word onsets with word number. The obtained approaches are promising and simpler than HMM-based methods. The results directly show that AF representations are as least as rich as manual segmentations on phone-level, and we argue that it is in fact a richer representation due to feature asynchrony. However, essentially different segmentations may result from other distance measures. To what extent metrics such as Kullback-Leibler show a similar performance is still unknown.

Other issues under investigation are the use of asynchrony for cue trading between AFs and the precise quantification of this asynchrony. The variation of observed AF vectors around a canonical AF vector is the combined contribution of both the feature asynchrony and the statistical variation in the classifier output, but this is still to be unravelled.

Finally, the observed relation between fine phonetic details and word type discrimination opens a challenging research area. AF representations provide an interpretable and rich representation which appears useful for research on lexical decoding and fine phonetic details.

6. Acknowledgements

Thanks to Mirjam Wester for assistance with the ANN set-up, and to Rob van Son for assistance in making the IFAcorpus available.

7. References

[1] Bael, C. van, Heuvel, H.v.d., Strik, H. (2004). Investigating Speech Style Specific Pronunciation Variation in Large Spoken Language Corpora. Proceedings of Interspeech (ICSLP) 2004, Jeju, Korea (cd-rom).

[2] Browman, C., Goldstein, L. (1992). Articulatory phonology: an overview. Phonetica 49, pp. 155-180.

[3] Chomsky, N., Halle, M. (1968). The Sound Pattern of English. Harper & Row, New York, NY.

[4] Frankel, J., Wester, M., King, S. (2004). Articulatory feature recognition using dynamic Bayesian networks. Proceedings of Interspeech (ICSLP) 2004, Jeju, Korea, (cd-rom).

[5] Hacioglu, K., Pellom, B., Ward, W. (2004). Parsing speech into articulatory events. In: Proceedings of ICASSP 04. Montreal (cd-rom).

[6] King, S., Taylor, P. (2000). Detection of phonological features in continuous speech using neural networks. Computer Speech and Language 14 (4), pp. 333-353.

[7] Li, J, and Lee, C.H.. (2005). On designing and evaluating speech event detectors. Proceedings Interspeech-2005 (cd-rom).

[8] Oostdijk, N. (2002). The design of the Spoken Dutch Corpus. In: Peters, P., Collins, P., Smith A. (Eds) New Frontiers of Corpus Research (pp. 105–112). Amsterdam: Rodopi.

[9] Ostendorf, M. (1999). Moving beyond the beads-on-a-string model of speech. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop. Vol. 1. Keystone, Colorado, USA, pp. 79-83.

[10] Richards, H. B., Bridle, J. S., (1999). The HDM: A segmental hidden dynamic model of coarticulation. In: Proceedings of ICASSP. Vol. 1. Phoenix, AZ, pp. 357-360.

[11] Strom, N. (1997). Phoneme probability estimation with dynamic sparsely connected artificial neural networks. The Free Speech Journal Issue 5.

[12] Wester, M. (2003). Syllable classification using articulatory-acoustic features. In: Proceedings of Eurospeech, Geneva, Switzerland (cd-rom).

[13] Wester, M., Frankel, J., King, S. (2004). Asynchronous articulatory feature recognition using dynamic Bayesian networks. In: Proceedings of the Institute of Electronics, Information and Communication Engineers Beyond HMM Workshop. Vol. 104. Kyoto, Japan, pp. 3742 (SP2004-81-95).

[14] Son, R.J.J.H. van, Binnenpoorte, D., Heuvel, H. van den, Pols, L. (2001). The IFA Corpus: a Phonemically Segmented Dutch "Open Source" Speech Database, Proceedings of Eurospeech, Aalborg, Denmark, Vol. 3, pp. 2051–2054.

[15] Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding, Journal of Phonetics, 31, pp. 373–405.

[16] Trubetzkoy, N. (1939). Grundzüge der Phonologie (Principles of Phonology). Travaux du Cercle linguistique de Prague 7.

[17] Wesenick, M.-B., Kipp, A. (1996). Estimating the quality of phonetic transcriptions and segmentations of speech signals. Proceedings ICSLP, Pittsburgh (cd-rom).

[18] Bosch, L. ten (2006). Speech variation and the use of distance metrics on the articulatory feature space. ITRW Workshop on Speech Recognition and Intrinsic Variation, Toulouse.

[19] Davis, M. H., Marslen-Wilson, W. D., Gaskell, M. G., 2002. Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition. Journal of Experimental Psychology: Human Perception and Performance 28, 218–244.

[20] Harrell, F., 2001. Regression modeling strategies. Springer, Berlin.

[21] Kemps, R., Ernestus, M., Schreuder, R., Baayen, R., 2005a. Prosodic cues for morphological complexity: The case of Dutch noun plurals. Memory and Cognition 33, 430–446.

[22] Kemps, R., Wurm, L., Ernestus, M., Schreuder, R., Baayen, R., 2005b. Prosodic cues for morphological complexity in Dutch and English. Language and Cognitive Processes 20, 43–73.

[23] Salverda, A., Dahan, D., McQueen, J., 2003. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition 90, 51–89.