

Towards a new level of annotation detail of multilingual speech corpora

Anja Geumann

Department of Computer Science
University College Dublin, Ireland
anja.geumann@ucd.ie

Abstract

The aim of this paper is to highlight the actual need for corpora that have been annotated based on acoustic information. The acoustic information should be coded in features or properties and is needed to inform further processing systems, i.e. to present a basis for a speech recognition system using linguistic information.

Feature annotation of existing corpora in combination with segmental annotation can provide a powerful training material for speech recognition systems, but will as well challenge the further processing of features to segments and syllables. We present here the theoretical preliminaries for our multilingual feature extraction system, that we are currently working on.

1. Introduction

A survey of speech corpora currently available (via *LDC* or *ELDA*) reveals a low amount of data that are phonetically annotated at a fine level of detail. This is of course not a surprise since the effort required to get fine-grained annotation is considerable, even at the segmental level as depicted in Figure 1. Automatic alignment could be useful and is used for a number of corpora (e.g. *Switchboard* and *VerbMobil*) but is used on a relatively broad level, but no feature-level annotation, automatic or manual is currently available.

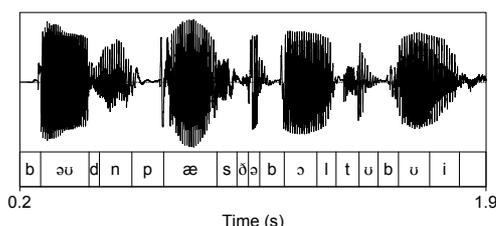


Figure 1: A segmental annotation, segments are manually annotated, based on auditory judgements.

The notion of different levels of labeling is of course not new and has been reviewed to some extent in [2]. However, what actually is on the market is disappointing.

We are currently developing a knowledge-based feature extraction system, to provide a way of enriching existing segmentally annotated corpora. This paper presents the theoretical preliminaries for our multilingual feature extraction system.

2. Features in speech recognition systems

Studies have been carried out, ([13] and references therein), that have compared the performance of phonological feature based systems as opposed to cepstral coefficients and conclude that they perform fairly similarly. The advantage of the use of features closer to phonological features is that they ease the burden on a further processing system, as they already anticipate some class building, yet remain closer to the acoustic signal, than a system based on segments. Indeed, there have been a number of feature-based speech recognition systems proposed, to name but a few see [1, 6, 7, 8, 13, 14, 15, 16]. While the feature classifications used in these systems exhibit similarities, they also differ in many details.

Acoustic correlates of features have been described in the literature (see especially [11, 18, 19]). A good overview and comparison of studies is given in [9]. The first detailed description of distinctive features [10] assumed that they had identifiable counterparts.

Feature-based speech recognition has to address two separate problems, the recognition of acoustic events and the mapping of a number of these to segments or higher-level units. By separating these two parts and focusing on the first, our aim is to substantiate the claims made in [9, 18, 19] and elsewhere that there are indeed sufficient correspondences between features, that have proven to be useful in an abstract phonological description, and parameters which are detectable from the signal via automatic means.

3. Multilingual perspective

In this section, we describe the concept of our multilingual feature inventory. It is based on experiences with other feature inventories see Figure 2 and [14, 15]. The feature annotation displayed in Figure 2 turns out to be insufficient for covering a variety of languages, although it might not be inherently language specific.

The motivation for the suggested feature inventory and the specification of sounds is based on phonological needs, i.e. to be able to distinguish between sounds of a language. However, for each feature and feature specification there must be a potential acoustic correlate. For example the feature *continuant* has a correlate in an abrupt change in the spectral pattern. Thus it seems not to be as easily applicable to nasals as to stops, so that nasals (see Table 1) are not specified for the feature *continuant* although they are in most phonological descriptions.

In order not to deviate too far from standard notation we have kept the names of phonological features including the standard counterparts, e.g. *voiced* - *voiceless*, *vocalic* -

nonvocalic. Segments have not to be specified for one of the counterparts. This leaves open the possibility of allowing the extraction of two apparently contradicting features as a way of representing a certain vagueness for this feature.

Our current goal is to model the inventory that needs to be detectable in six languages, namely English, French, German, Irish, Romanian and Spanish. The inventory used in [22], and shortly described in Figure 2 is not adequate to model all six languages as it excludes descriptions of secondary articulation as found in Irish, e.g. velarization [m^v] and palatalization [m^j] or nasalization of vowels as in French.

We assume that the features are not language specific. Thus the restriction to a small number of tiers where features as *nasal* and *vocalic* compete proves as inadequate (Figure 2) for the description of any language.

The feature specification of sounds should stay as close to a phonological description as possible since this is for many languages at best the only available information.

Even a feature as voicing should be described on two separate tiers, first to avoid a notion of binarity and second to allow *voiced* and *voiceless* to be interpreted as equally notable.

The proposed feature set may have to be extended for the description of more languages. Currently, voice quality is described in fairly broad terms with respect to the features *voiced*, *voiceless*, *pos-VOT*, *neg-VOT*. However, the extensions should not have an impact on the overall architecture, but be interpreted more or less as additions.

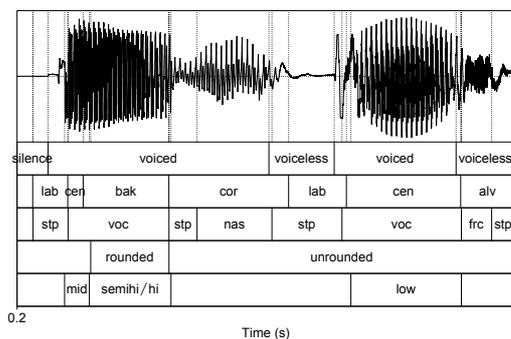


Figure 2: A feature annotation adapted in [22] from [6] and [14]. Features are placed on five tiers: voicing, place, manner, rounding, vowel height. Features for this example are manually annotated, based on auditory judgements.

4. Further considerations

A number of issues remain to be addressed in connection with the so far proposed feature-based annotation. They are pointed out briefly here.

4.1. Complex segments, duration

Currently, in the inventory, complex segments (affricates, diphthongs) are described as a sequence of separate sounds. This might have to be modified in the future.

The distinction between Spanish [r] and [r̄] is described as a pure length distinction. Following the definition of acoustic events [4], features are assumed to have a temporal extension, allowing temporal overlap as well as length information to be further processed. Thus languages that exhibit genuine length distinctions rely on the temporal extension of the features associated with it.

4.2. Feature organization, interaction

Although it was stated before that the feature specification of segments is initially very much oriented on phonological descriptions of the respective language, we assume no feature hierarchy or ordering. The organization of features is interpreted as a *posteriori* structuring. This is discussed in detail in [17, 5].

It might turn out to be useful to value some features higher as others, similarly to the *landmarks* by Stevens [19].

4.3. Mapping from features to segments

An open issue in the use of features whether they are cepstral coefficients or higher level features is how to interpret their asynchronicity and map them onto segments or syllables. We believe that this crucial open issue can be much easier addressed once there are a number of data available that provide feature information.

The amount of features is probably best handled in a system as proposed in [22], which is based on [4], and will allow for errors in the input by using higher level linguistic information for reasoning.

4.4. Evaluation

The evaluation of feature annotations is of course not trivial. As stated above no manual annotation of a quantity worth mentioning exists for the purpose of comparison. It is more likely that feature annotations will be evaluated with respect to patterns, which have been derived from detailed segmental descriptions [6], see Figure 1. Using a lookup table this segmental annotation could be expanded to a bundle of simultaneous features, which should hypothetically resemble the detected features in their temporal extension.

4.5. Extensibility

It might prove useful to extract further spectral information and add this uncategorised to the feature classification. The overall intensity for example could be useful for the prosodic interpretation.

Additionally, no thought on representations for tone languages has been spent, which is a clear gap that has to be filled rather sooner than later.

Taken all these parameters into consideration we can come up with the following Table 1, as extract of a list of 88 segments (60 consonants, 28 vowels), describing the inventories of the six languages mentioned. The features we are here suggesting are not abstract, but have acoustic correlates, which have been reported on elsewhere, e.g. [8, 9, 18, 19].

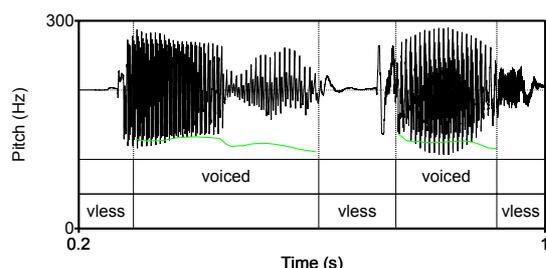


Figure 3: An extract of proposed feature annotation. Features are located on separate tiers. Features are extracted automatically, here equivalent to pitch detection, using PRAAT [3] and written to separate annotation tiers.

5. Conclusions and Outlook

We feel by the recently refreshed interest in phonological features (sometimes referring to IPA classification and called articulatory features, e.g. [20]) in speech recognition systems supported in our view that this is the most promising way speech recognition should go. However the certain amount of arbitrariness in the used inventories and their organization on tiers or not, is so far confusing. The features we are here suggesting are not abstract, but have acoustic correlates, which have been reported on elsewhere, e.g. [8, 9, 14, 15]. We would like to see this as a proposal towards establishing a set of standards for feature annotation, similar to segmental annotation standards as the use of the International Phonetic Alphabet, where symbols have a defined meaning and the set of symbols is defined.

6. Acknowledgements

The author would like to thank Julie Berndsen and Moritz Neugebauer for lively discussions and comments on this paper.

This material is based upon works supported by the Science Foundation Ireland under grant No. 02/INI/II00. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of SFI.

7. References

- [1] Ali, A.M.A., van der Spiegel, J., Mueller, P., Haentjens, G., Berman, J. "An acoustic-phonetic feature-based system for automatic phoneme recognition in continuous speech". *IEEE international symposium on circuits and systems, ISCAS-99*, vol 3:118-121, 1999.
- [2] Barry, W. J and Fourcin, A. J. "Levels of Labelling", *Computer, Speech and Language* 6:1-14, 1992.
- [3] Boersma, P., Weenink, D. PRAAT: doing phonetics by computer. Institute of Phonetic Sciences, Amsterdam [<http://www.praat.org>].
- [4] Carson-Berndsen, J. *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition*. Dordrecht, Holland: Kluwer Academic Publishers, 1998.
- [5] Carson-Berndsen, J., Geumann, A., Neugebauer, M. "Embracing Multilinguality: Defining Phonetic Features for Speech Technology" Article in prep.
- [6] Chang, S., Greenberg, S. and Wester, M. "An Elitist approach to articulatory-acoustic feature classification", *Proc. 7th Eurospeech*, Aalborg, Denmark, 1725-1728, 2001.
- [7] Espy-Wilson, C. "A feature-based semivowel recognition system", *JASA* 96(1):65-72, 1994.
- [8] Espy-Wilson, C. "Acoustic measures for linguistic features distinguishing the semivowels /w j r l/ in American English", *JASA* 92(2):736-757, 1992.
- [9] Harrington, J. "Acoustic cues for automatic recognition of English consonants", in: M. Jack and J. Laver (eds.) *Speech technology: A survey*. Edinburgh: Edinburgh University Press, 69-143, 1988.
- [10] Jakobson, R., Fant, G., Halle, M. *Preliminaries to speech analysis: The distinctive features and their correlates*. MIT Press, 9th ed. 1969 (1952).
- [11] Juneja, A., Espy-Wilson, C. "An event-based Acoustic-phonetic approach for speech segmentation and e-set recognition". Proc. of 15th ICPHS, Barcelona, 2003.
- [12] Juneja, A., Espy-Wilson, C. "Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning". Proceedings of ICNIP 2002.
- [13] King, S., Taylor, P., Frankel, J. and Richmond, K. "Speech recognition via phonetically-featured syllables", Saarbrücken: Institute of Phonetics WP, University of the Saarland. *PHONUS* 5:15-34, 2000.
- [14] Kirchhoff, K. "Integrating articulatory features into acoustic models for speech recognition", Saarbrücken: Institute of Phonetics WP, University of the Saarland. *PHONUS* 5:73-86, 2000.
- [15] Koreman, J. and Andreeva, B. "Can we use the linguistic information in the signal?" Saarbrücken: Institute of Phonetics WP, University of the Saarland. *PHONUS* 5: 47-58, 2000.
- [16] Lahiri, A. and Reetz, H. "Underspecified recognition", in: C. Gussenhoven and N. Warner (eds.), *Laboratory Phonology 7*. Mouton de Gruyter: Berlin, New York, 637-675, 2002.
- [17] Neugebauer, M. "Computational Phonology and Typed Feature Structures". Proc. of the First CamLing Postgraduate Conference on Language Research, University of Cambridge, 2003.
- [18] Stevens, K.N. "Acoustic correlates of some phonetic categories", *JASA* 68(3):836-842, 1980.
- [19] Stevens, K.N. *Acoustic Phonetics*, MIT Press: Cambridge (Ma), London, 1998.
- [20] Stüker, S., Schultz, T., Metze, F., Waibel, A. "Multilingual Articulatory Features". *Proceedings of ICASSP 2003*, vol. I:144-147.
- [21] Sun, J. and Deng, L. "An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition". *JASA* 111(2):1086-1101, 2002.
- [22] Walsh, M. "Recasting the time map model as a multi-agent system". Proc. of the 15th ICPHS, Barcelona:735-738.