Exploring phonetic realization in Danish by Transformation-Based Learning

Marcus Uneson, Ruben Schachtenhaufen***

Lund University, Copenhagen Business School***

Abstract

We align phonemic and semi-narrow phonetic transcriptions in the DanPASS corpus and extend the phonemic description with sound classes and with traditional phonetic features. From this representation, we induce rules for phonetic realization by Transformation-Based Learning (TBL). The rules thus learned are classified according to relevance and qualitatively evaluated.

Introduction

Language abounds with classification tasks – some we solve ourselves, some we hand over to machines. In the latter case, we may or may not be interested in what the machine actually learns. Stochastic classifiers such as HMMs and SVMs are useful for many purposes, but their target representation is usually inscrutable to humans. Rule learners, on the other hand, may or may not match stochastic classification performance, but what they learn may be interesting in itself – sometimes, it might even be the main point.

In the present paper, we explore one of those cases: the application of a well-known rule induction technique, Transformation-Based Learning (Brill, 1995), on phonetic string representations. The problem can be phrased thus: given a phonemic and a semi-narrow phonetic transcription of speech, can we extract transformation rules which will take the first to the second, or at least part of the way? If so, do these rules give us any new insights? Somewhat less abstractly, our aim is to automatically induce typical textbook rules for phonetic realization, from a transcribed, real-world corpus of spontaneous, connected speech. The language under study is Danish, where, arguably, the distance between these two representations is particularly noteworthy.

Background

Danish phonology

Grønnum (2005) analyzes Danish phonology into 11 vowel phonemes (/i e ɛ a y ø œ u o ɔ ə/) and 15 consonant phonemes (/m n p t k b d g f v s h l r j/), plus the prosodic elements *stød*, length, and stress. Briefly, most non-high vowels are realized more open before and/or after /r/, and some consonants are realized differently depending on syllable position: /p t k/ are aspirated in onset and unaspirated in coda; /d g v r/ are contoid in onset and vocoid or ∅ in coda.

The realization of /ə/ is quite complex. More often than not it is elided, leaving its syllabic trait and compensatory lengthening on adjacent sounds. The combination of consonant gradation in coda and a very fleeting /ə/ results in a highly unstable sound structure in current Danish.

In traditional descriptions, being based mainly on conservative, careful, read speech, Danish phonemes typically have one or two, rarely three, allophones, e.g. "/d/ > [ð] in coda, [d] elsewhere". In spontaneous speech phonemes have a much wider range of realization – for instance, in DanPASS (see below), /d/ is transcribed [d ð ɾ ɹ t z s], among others.

Transformation-based learning

Transformation-based learning (TBL) was proposed by Eric Brill (Brill, 1995). It is, in a one-sentence summary, a supervised machine learning method producing a compact, ordered, human-readable list of classification rules (or *transformations*), each chosen greedily from a set of candidates dynamically calculated from user-supplied patterns (the *templates*), so that it maximally reduces (a function of) the difference between the system's present idea of the classification (the *current corpus*) and a gold standard (the *truth*). One sentence is likely not enough; we refer to Brill (1995).

The task at hand reminds somewhat of letter-to-sound (LTS) conversion, to which TBL also has been applied (Bouma, 2000). Abstractly, both problems concern transforming one string representation of language into another.One major difference is that LTS aims at lexical pronunciation:

	der igen	er i	midten
phonemic:	deːʔr_iˈgɛn	ɛr_iːʔ	ˈmetən
phonetic:	daɪˈgɛn	ˈaɪ	ˈmedn̩

Figure 1: DanPASS phonemic and phonetic tiers for der igen er i midten *'that again is in the middle'*

it usually has a well-defined target. Phonetic realizations, by contrast, have several influencing factors but few truly functional dependencies. In this paper we will pay more attention to the rules themselves extracted than how close to the (partly arbitrary) target they will take us.

The present study

On transcriptions

Although historically much used, the method of taking transcriptions as point of departure for phonetic conclusions is not without its problems. Transcriptions imply a simplistic and much reduced 'beads-on-a-string' view on speech, often with weak support in data which have not been filtered through the perception of a native speaker. In the words of Grønnum (2009), "phonetic notation, specifically of the rather narrow kind, and prosodic labeling are both impressionistic exercises". For the purposes of this paper, however, we will accept this armchair view.

The DanPASS corpus

Our data is certainly not armchair; it was taken from the DanPASS corpus[1] (Danish Phonetically Annotated Spontaneous Speech) (Grønnum, 2009). In total, the corpus comprises about 10 hours (73kW) of annotated high-quality recordings of connected speech produced by 27 speakers, distributed among several tasks in non-scripted monologue and dialogue. DanPASS addresses no particular research need specifically, but is generally well suited for studying phenomena associated with connected, spontaneous speech. With the exception of a small fraction of non-spontaneous speech (elicited word lists), we used all of it.

The phonemic transcription in DanPASS is based on the analysis of Grønnum (2005) mentioned above. The annotations of DanPASS are available as Praat tiers. The ones of concern here are the the phonemic notation and and the semi-narrow phonetic notation. Figure 1 shows a small corpus sample for these.

[1] http://danpass.dk

Experimental setup

We employed the μ-TBL system (Lager, 1999), with the semi-narrow transcription tier taken as truth and the phonemic tier as the initial current corpus. TBL requires that the current corpus and the truth are containers of the same shape, which in the present case requires alignment of the transcriptions; for this task, we used the sound class alignment method proposed by List (2010).[2] Since our interest lies in rules which apply with few or no exceptions, all rules were required have a minimum accuracy of 0.95.

The problem encoding required more consideration. Rules should of course be conditioned on the immediate phonetic context. Importantly, however, the learner should also be capable of at least simple generalizations: if rule R applies in phonetic environment A, and phonetic environment B is "similar" to A, then maybe R applies in B as well? One way to operationalize the notion of similarity is to partition the phoneme set into predefined sound classes; another is to allow subphonemic descriptions. On a closer look, these are actually not very different: they both define characteristic functions and allow a rule learner to construct predicates on a given environment.

In the experiments described below, the sound classes follow a suggestion by Dolgopolsky, as adapted and extended by List (2010). In principle, the entire IPA space is partitioned into the classes in Table 1. The subphonemic description of a phoneme is simply its associated features in the traditional sense, treated as sets. A sample of the corpus thus encoded (which also exemplifies the alignment) is given in Table 2.[3]

The TBL templates chosen to operate on these features are given in condensed form below. An additional constraint was that an elided segment would not be subject to further changes.

Change segment A to segment B when …

- …(left/right) (segment/segment class) is X;
- …left (segment/segment class) is X and right (segment/segment class) is Y;
- …(left/right) (segment/segment class) is X and the next neighbour (segment/segment class) is Y;
- …(left/right) segment has feature F;

[2] http://lingulist.de/lingpy/

[3] It is worth noting that μ-TBL permits Prolog code as part of the template specifications, thus forming a little embedded language. Whether the class and phonetic features of Table 2 were prespecified or calculated dynamically only influences running time and memory use, not semantics. This is very useful in interactive experimentation.

Table 1: Dolgopolsky sound classes for phonological rules, as adapted by List (2010)

Code	Segment	Example
P	labial obstruents	p,b,f
T	dental obstruents	d,t,θ,ð
S	sibilants	s,z,ʃ,ʒ
K	velar obstr.; dent. & alv. affricates	k,g,ts,tʃ
M	labial nasal	m
N	remaining nasals	n,ɲ,ŋ
R	liquids	r,l
W	voiced labial fric.; init. rd. vowels	v,u
J	palatal approximant	j
H	laryngeals and initial velar nasal	h,ɦ,ŋ
A	all vowels	a,e,i

- …current segment shares feature F with (left/right/left and right) segment.

Results and discussion

From a training material of 210,000 phonemes and at score threshold of 10, the system learned 446 rules in about six hours. On unseen test data, the learned rules took the correspondence between truth and hypothesis from 41.2% to 73.2%. Neither of these numbers is very informative: the TBL evaluation function assumes a unique notion of truth, but for a given phonemic representation, there are many acceptable phonetic realizations. Even more ambiguously, for two given strings, one phonemic and one phonetic, there are many reasonable rule sequences that transforms the first into the second.

Indeed, quantitative evaluation of the learned rules is a challenge. One possibility is to arrange perception tests on the naturalness of the generated pronunciations; but this says nothing about the phonological validity of individual rules. For this paper, we opted for a less formal, manual evaluation. For coverage, we contented ourselves with noting that at a glance, the rule list appears to contain the majority of allophonic alternations in Danish. For rule accuracy, the main interest here, we took the first 108 rules (those with score > 100) and classified them as follows (ordered according to our intuitive idea of rule "quality"):

1. (3) False in Danish, unexplainable in data
2. (27) False in Danish, attributable to data
3. (74) Largely in agreement with current descriptions of Danish:
   (a) (49) inaccurate, could be more refined
   (b) (15) true, satisfyingly general
   (c) (10) true, interdependent with other rule/s found
4. (4) Interesting: not in agreement with current descriptions but possibly a new phonological development in progress

Table 2: The DanPASS sample of Figure 1, where phonemes are encoded with their identity (phm), class (cls), and features. Implicit time axis runs from top to bottom. The two left columns also show the resulting alignment of the phonetic (pht) and phonemic transcription.

Pht	Phm	Cls	Features
d	d	T	['voiced', 'alveolar', 'plosive']
aɪ	eːʔ	A	['length-mark', 'plosive', 'glottal', 'front', 'unrounded', 'close-mid']
-	r	R	['voiced', 'alveolar', 'trill']
-	i	A	['front', 'close', 'unrounded']
g	g	K	['voiced', 'velar', 'plosive']
ɛ	ɛ	A	['front', 'open-mid', 'unrounded']
n	n	N	['alveolar', 'nasal']
aɪ	ɛ	A	['front', 'open-mid', 'unrounded']
-	r	R	['voiced', 'alveolar', 'trill']
-	iːʔ	A	['length-mark', 'plosive', 'glottal', 'unrounded', 'front', 'close']
m	m	M	['nasal', 'bilabial']
e	e	A	['front', 'unrounded', 'close-mid']
d	t	T	['voiceless', 'alveolar', 'plosive']
-	ə	A	['schwa']
n̩	n	N	['alveolar', 'nasal']

Table 3 gives a few induced rules, chosen for illustration of the categories listed. In the following comments, "C#m" refers to the categories in the list above and "R#n" to the leftmost column of Table 3 (i.e., the position of the rule in the learned sequence).

Three of the learned rules make no sense, neither for Danish in general nor for DanPASS (C#1, R#93). These are as far as we can tell artefacts of the combined tokenization–alignment process.

More interestingly, several rules are found which are false for Danish but can be said to be true for the data (C#2). Such rules can be attributed to reductions which are uncommon in types but common in tokens (occurring, say, in a few, high-frequent function words). For instance, R#14 emanates from the modal verb /skal/ *skal* 'shall, must'. Usually, this is reduced to [sga].

The majority of the rules (C#3) can be described as reasonable, but not very interesting (outside verifying the validity of the procedure). Many of them are overly specific and would gain

Table 3: Some induced rules, in μ-TBL syntax. For instance, pht:A>B ← class:'C'@[-1] & feature:'F'@[1] *means that A transforms to B when the previous segment ([-1]) belongs to class C (Table 1) and the following ([1]) has feature F*

#	Score	Rule
1	11437	pht:r>∅ ← class:'A'@[-1]
5	1569	pht:ə>∅ ← class:'N'@[1]
14	752	pht:l>∅ ← phm:a@[-1] & phm:k@[-2]
18	605	pht:ə>ɐ ← class:'K'@[-1] & phm:r@[1]
20	561	pht:g>∅ ← phm:ə@[1]
24	458	pht:k>ɣ ← phm:ə@[1] & phm:r@[2]
29	384	pht:ə>ɐ ← class:'W'@[-1] & class:'R'@[1]
74	143	pht:ʁʔ>ʔ ← feat:open@[-1]
93	112	pht:ɒʔ>ɔʔ ← feat:approxim@[1]

from generalization (C#3a, Rs#18,29). However, some generalizations (C#3b, R#74) are indeed discovered. As is typical to phonology, many rules have a feeding order and can only be evaluated in conjunction with other rules found. In most cases the system finds such interdependent rules (but does not connect them) (C#3c, R#5).

Finally, some genuinely interesting rules are also discovered (C#4) that might for instance indicate ongoing phonological change. Thus, Rs#24,1,20) suggest progressive consonant lenition or elision, postvocalic or pre-schwa.

Conclusion

This paper presents an attempt to extract phonetic realization rules from transcribed spontaneous speech, by conditioning on local phonemic context only. Of course, we recognize that this is insufficient for real-world data, where phonetic variation can only partly be described by phonology. Other extra-phonological (information structure, word frequency, etc) and extralinguistic (speaker style, speaker mood, speech rate, acoustic environment, etc) factors are equally important. We also recognize the difficulties in evaluation, and the more general problems associated with doing phonetics on transcriptions. Nevertheless, for a first attempt, we find the results interesting, at least enough to pursue further.

One obvious source of potential improvement is additional features describing the phonetic and linguistic environment. Some of the relevant linguistic factors are readily available for featurization. DanPASS already has annotations of basic information structure and part-of-speech. At present, syllable boundaries are not part of the DanPASS phonemic annotation tier, but the second author is currently preparing their inclusion.

As mentioned, some inappropriate rules can be attributed to reductions appearing in a few, high-frequent words. Clearly, 10000 occurrences of a certain reduction in a single, high-frequent word carry much less phonological evidence than 100 occurrences in each of 100 different words. This observation could be exploited; e.g, by binning phonetic environments into lexical contexts and weighting those contexts sublinearly (e.g., logarithmically), much as sublinear term frequency scaling is used in information retrieval.

A more general problem is that of undiscovered rule generalizations. Although the current system can examine phonetic features of its surroundings, the rules work at phoneme level only. A more fine-grained representation might be beneficial, where rules are allowed to add or remove individual phonetic features. This would allow generalizations such as ”add voice to voiceless stop between two vowels”. Again, however, the more fine-grained the representation, the more fragile the beads-on-a-string assumption, and the higher the number of competing notions of truth.

Adding expressivity to the horizontal rather than the vertical direction, one might let the system simultaneously replace more than one segment. This is not very interesting for general TBL, as it comes with the cost of a much expanded search space and buys little or nothing in performance. However, in the present task the alphabet is small and the rules are the target, and it might be worth the effort.

References

Bouma G (2000). A finite state and data oriented method for grapheme to phoneme conversion. In *NAACL-2000*, 303–310. Seattle, WA.

Brill E (1995). Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21(4):543–565.

Grønnum N (2005). *Fonetik og Fonologi – Almen og Dansk*. Akademisk Forlag, Copenhagen.

Grønnum N (2009). A Danish phonetically annotated spontaneous speech corpus (danpass). *Speech Communication*, 51:594–603.

Lager T (1999). The μ-tbl system: Logic programming tools for transformation-based learning. In *Proceedings of CoNLL*, vol. 99.

List J M (2010). Phonetic alignment based on sound classes. In M Slavkovik, ed., *Proceedings of ESSLLI 2010, Student session*, 192–202.