Acoustic word disambiguation with phonological features in Danish ASR

Andreas Søeborg Kirkedal
Interactions, LLC
Murray Hill, NJ, USA
akirkedal@interactions.com

Abstract

Phonological features can indicate word class and we can use word class information to disambiguate both homophones and homographs in automatic speech recognition (ASR). We show Danish stød can be predicted from speech and used to improve ASR. We discover which acoustic features contain the signal of stød, how to use these features to predict stød and how we can make use of stød and stød-predictive acoustic features to improve overall ASR accuracy and decoding speed. In the process, we discover acoustic features that are novel to the phonetic characterisation of stød.

1 Introduction

Stød ([ˀ] in IPA notation) is usually described as (a kind of) creaky voice or as *laryngealisation* (Hansen, 2015; Grønnum et al., 2013). Stød can distinguish homophones and homographs and can identify word class by its presence. Danish *viˀser* is a noun that translates to *clock dial*, but pronounced without stød - *viser* - it can also be a verb that means *to show*. The presence of stød can change the meaning of an utterance, e.g. *de kendte folk* can mean *the famous people* if *kendte* is pronounced as [kɛnˀdə] and can also mean *they knew people* if *kendte* is pronounced as [kɛndə]. Stød is robust against some types of reduction and is an acoustic cue that can help distinguish *one* vs. *none* in colloquial Danish: [eˀn] and [eŋ].

Phonological features can often be determined from grammar and morphology (Grønnum, 2005) but stød may not occur in read-aloud or spontaneous speech when predicted by morphology and grammar, and stød can be difficult to perceive in both visualisations of spectrograms and in speech (Hansen, 2015).

Stød is not highly frequent in either read-aloud or spontaneous speech, but stød and similar phonological features like e.g. tones are interesting for two main reasons:

1. Relatively small languages like Nordic languages do not have large speech corpora available like English, Chinese etc. We should exploit all signals in the data to improve ASR performance for these languages.

2. The semantic disambiguation at both sentence and lexical level is appealing because ASR errors that disturb the meaning of an utterance are less acceptable for human consumers of ASR output (Mishra et al., 2011).

Our contributions are to:

- Show that stød annotation is reliable when annotated by trained phoneticians and can be the basis of statistical analyses.
- Discover novel audio features that are predictive of stød in speech.
- Demonstrate we can predict stød in speech as phone variant discrimination.
- Integrate stød in ASR and improve WER on read-aloud and spontaneous speech.

2 Related work

Henrichsen and Christiansen (2012) found a correlation between fundamental frequency (F0) and spectral tilt, and discrimination between *content* words and *function* words. We also investigate these features for their ability to predict stød.

Stød, stress and schwa-assimilation were studied in Kirkedal (2013) in an ASR context. The study found that WER improved when stød, stress, schwa and duration annotation were removed from the lexicon. However, the ASR system was trained on a single speaker corpus with read-aloud speech

Proceedings of the 15th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 21–31

https://doi.org/10.18653/v1/P17

Dataset	Speech genre	Location	Speakers	Duration	Types	Tokens
Språkbanken-train	Read	Office	560	316h	65667	2366183
Språkbanken-test	Read	Office	56	77h	72978	185049
JHP	Spontaneous	High-school	2	1 min	178†	995†
PAROLE48	Read	Speech lab	1	48 min	181	4705
DanPASS	Spontaneous	Speech lab	18	2h 51 min	1075	21170

Table 1: Summary table for the corpora used. † indicate that type/token counts are based on counts of phones instead of words.

and evaluated on a test set from the same corpus and we demonstrate that the findings do not generalise to a multi-speaker setting.

No single feature extracted from audio of spoken Danish can predict the presence of stød like F0 estimation can predict pitch (Fischer-Jørgensen, 1989). Because stød is related to irregular vibration of the vocal folds, previous research has focused on harmonics-to-noise (HNR) ratio, the difference between the first two harmonics in a spectrum (H1:H2) and diplophony (H1:H1$\frac{1}{2}$[1]) as well as F0 and intensity (Hansen, 2015), but this is the first large scale quantitative study of stød.

Stød can be audibly heard yet not be visible in a spectrogram to an experienced researcher (Hansen, 2015). Consequently, the annotation of stød is subject to annotator perception. Annotators need a considerable amount of training to be able to annotate stød and the high cost of annotation in terms of training and annotation time coupled with potential bias from annotator training or the specific annotator has been a barrier to quantitative studies of stød. We show that expert stød is reliable in Section 4.

Like stød in Danish, Tone 1 and Tone 2 in Norwegian and Swedish are the only difference between some homographs and homophones. Swedish and Norwegian are pitch accent languages that use tones to distinguish lexical items that would otherwise be homophones and homographs, e.g. *tanken*$_1$ vs. *tanken*$_2$ (*the tank* vs. *the thought* - subscript indicates Tone 1 and Tone 2) (Lahiri et al., 2005). Some theories suggest that stød originated from tones and the distribution of stød and Tone 1 & 2 also show similarities (Grønnum et al., 2013). Riad (2000) describe stød as a tonal pattern but this is refuted in a reply in Grønnum et al. (2013).

In tonal languages like Mandarin Chinese, tones or tonal contours disambiguate monosyllabic words as in the famous example of *ma* which has five different meanings depending on the tonal contour. ASR for tonal languages add suprasegmental information to ASR models either by extending the acoustic feature input (*embedded*) or rescoring word lattices (*explicit*) (wen Li et al., 2011). Embedded modelling requires that tones are modelled in the lexicon either as tonal variations of the same phoneme (Metze et al., 2013; Yoon et al., 2006) or as separate phonemes (Adams et al., 2018). Stød is related to irregular vibration of the vocal folds which occurs frequently in Danish with no connection to stød and we do not explore explicit modelling.

The duration of the stød-bearing (semi-)vowel or syllable has been considered important in previous literature. We do not consider duration in this paper for 2 reasons: 1) HMM-based ASR is the target application and implicitly model duration with self-loops in the HMM and 2) the investigations of duration where conducted in lab conditions with elicited speech in the Standard Copenhagen dialect. We use several corpora that cover most Danish dialects, also dialects that typically do not use stød.

The rest of the paper is structured as follows: Section 3 presents the data used and Section 4 presents the study of stød annotation. In Section 5 we discover novel acoustic features that are predictive of stød. We test and evaluate how well acoustic features predict stød in Section 6 and perform phone variant discrimination where we jointly predict phone and stød. In Section 7, we adapt an ASR recipe for Danish and train several ASR systems to determine the best way to use stød to improve ASR.

3 Data

Table 1 shows the corpus statistics for all corpora used in the rest of this paper.

[1]The ratio between the first harmonic H1 and the harmonic signal at F0 $*$ 1.5. F0 is the frequency of H1.

We use an interview with a high school student in real-world conditions, denoted as JHP[2] to study the reliability of stød annotation. We use this short sample of speech because it is the only sample that is annotated by four Danish-speaking expert phoneticians trained in stød annotation. Another expert phonetician aligned and time-coded the four transcriptions.

We also use the monologues from DanPASS (Grønnum, 2006) and speech from PAROLE-DK[3] (Henrichsen, 2007). To compensate for the unequal corpus sizes, we sample only 48 minutes and refer to this subset as PAROLE48. We separate a random subset that contains speech from both DanPASS and PAROLE48 as a test set.

Nasjonalbiblioteket[4] hosts a language repository called Språkbanken. In the repository is a multilingual speech corpus also known as Språkbanken. The Danish 16 kHz part of Språkbanken contains recordings of phonetically-balanced utterances and covers 7 regions of Denmark and ages ranging from 18-70. The Swedish part was used in Vanhainen and Salvi (2014) to create an ASR recipe.

Språkbanken-test is 15 times larger than standard test sets from the Linguistic Data Consortium (LDC) such as HUB5 which is 5 hours long.[5] We decided to split Språkbanken-test into a development set SPDEV (ca. 9 hours) and test set SPTEST (ca. 17 hours). The remaining 51 hours are included in the training data (ca. 367 hours) while making sure that neither speakers nor utterances in SPDEV and SPTEST appear in the training set.

We create pronunciation lexicons with eSpeak (Duddington, 2010) from the training transcripts because the pronunciation lexicon distributed with Språkbanken has low coverage and eSpeak was found to produce transcriptions that are good enough for ASR (Kirkedal, 2014).

4 Stød annotation study

The data we use for training and testing needs to be reliable, i.e. if stød is annotated, we need to be sure that stød occurs. To test how reliable our data is, we calculate inter-annotator agreement measured by Cohen's κ and an *annotator competence score* (ACS) with MACE (Hovy et al., 2013). ACS is based on an item-response model that assumes an annotator will produce the correct phone sequence if he tries to which is valid in this scenario. An *item* is a unit in the phone sequence and in this study each unit is labelled by 4 annotators. We use both κ and ACS because κ is a measure of the annotation whereas ACS is an estimate of annotator proficiency. For both κ and ACS, higher scores are better. 7.8% of the items in JHP are annotated with phones with stød (stød-bearing) and the phones without stød (stød-less) will dominate κ because the distribution of phones in JHP is Zipfian and all stød-bearing phones are in the long tail. To focus specifically on stød, we report κ computed over stød-bearing items in two conditions:

1. Items that are labelled with stød by at least one annotator e.g. [ð$^{?}$], [ɑ$^{?}$], [n$^{?}$] etc.

2. The same items as in 1. but binarised such that e.g. [ð$^{?}$], [ɑ$^{?}$], [n$^{?}$] → 1 and [ð], [ɑ], [n] → 0.

We compute ACS over all phones and over the binarised stød annotation in 2.

We discovered 10 errors in the data, e.g. one label was [$^{?}$n], but should have been [n] and stød should have been annotated on the previous phone as [ð$^{?}$]. There were 7 alignment errors that was caused by the interpretation of a syllable nucleus as either two short vowels or a long vowel. This has an impact on the alignment because stød is annotated on a syllable rather than a phone and the data is aligned at the phone level.

We corrected the errors before calculating the κ scores based on phones and binary stød in Table 2. Average κ is a an average over all pairwise κ scores where the specific annotator is involved. The annotators are referred to as IPA1, IPA2, IPA3 and IPA4.

The κ scores in Table 2 and the ACS scores in Table 3 both indicate that stød annotation is reliable and we can base statistical models on stød

	IPA1	IPA2	IPA3	IPA4
Phone avg.	0.82	0.80	0.81	0.85
Stød avg.	0.72	0.74	0.76	0.76

Table 2: Average κ inter-annotator agreement on stød-bearing items.

[2]The JHP sample was made available by Jan Heegård Petersen, Copenhagen University.

[3]This corpus was used in Kirkedal (2013).

[4]The Norwegian National Library service.

[5]See https://catalog.ldc.upenn.edu/LDC2002S13.

Annotator	# labels	Phone	# stød	Stød
IPA1	107	0.760	53	0.770
IPA2	99	0.813	58	0.840
IPA3	94	0.823	62	0.894
IPA4	107	0.833	59	0.856

Table 3: Annotator competence scores for all items and stød-bearing items.

annotation. The high κ scores show that the annotation in JHP is high quality and the ACS scores show that the annotators are able to annotate stød consistently and accurately.

5 Acoustic correlates of stød

Because we can rely on expert stød annotation, we can discover acoustic features that signal stød with statistical models. We use DanPASS and PAROLE48 as training and test data because they are also manually annotated and there is annotator overlap with JHP. We use the toolkits Kaldi (Povey et al., 2011), Covarep (Degottex et al., 2014) and Praat (Boersma, 2002) to extract features that may contain information that signals the occurrence of stød. The number of features extracted by the different toolkits can be seen in Table 4.

We sample the audio every 10 milliseconds and extract features over a context window the size which depends on the feature. Mel-feature cepstral coefficients (MFCC) and perceptual linear prediction (PLP) features use a 25 ms window while pitch estimation uses a 1.5 second window to extract robust features. Each 10 ms, we extract MFCC features, PLP features, Phase Distortion Mean (PDM) features, Phase Distortion Deviation (PDD) features, the Maxima Dispersion Quotient (MDQ), Peak slope (PS), Quasi-Open Quotient (QOQ), Normalised Amplitude Quotient (NAQ), Parabolic Spectral Parameter (PSP), the difference between first and second harmonic (H1-H2), Fant's basic shape parameter (R_d[6]), HNR and Intensity[7]. The first coefficient (C0) is replaced by an energy feature in both MFCC and PLP extraction and we choose to discard the energy feature from MFCC extraction and keep the log-energy feature with derivatives from PLP extraction. When referring to 1st and 2nd derivatives, we will suffix the feature name with *-d* and *-dd*, respectively. Δ-pitch is a derivative on the raw unnormalised pitch estimate in log space computed over 5 frames and log-pitch is mean subtracted by an average pitch value over a 151 frame context window that is weighted by a probability of voicing feature (PoV). We also estimate pitch with Praat because Praat and Kaldi behave differently in unvoiced speech: Kaldi interpolates the pitch estimate across unvoiced regions and Praat sets it to 0.

[6]See (Fant, 1995) for a description.

[7]We use amplitude and intensity interchangeably, but we are aware that amplitude is the acoustic correlate of intensity.

Toolkit	**Dimension**	**Feature**
Kaldi	39	PLP*
	3	PoV*
	3	log-pitch*
	3	Δ-pitch*
Covarep	24	MFCC/MCEP
	25	PDM
	13	PDD
	1	MDQ
	1	PS
	1	QOQ
	1	NAQ
	1	PSP
	1	H1-H2
	1	R_d
Praat	1	HNR
	1	Intensity
	1	Pitch

Table 4: Acoustic features. Features marked with * also include 1st order and 2nd order derivatives.

We align each 120-dimensional feature vector to a single phone by first segmenting syllable and word level annotation to phone level and relying on the existing time-coding.

5.1 Ranking acoustic features

We want to rank the 120 features by how well they predict stød with Extremely-Randomised Trees (Geurts et al., 2006) which trains an ensemble of decision trees. Decision trees can use features without standardisation and the input is not assumed to be normally distributed, which is not the case for e.g. HNR which becomes undefined if the harmonic component of the speech signal becomes too noisy. The estimation of relative feature importance will also be less affected by differences between the toolkits.

The algorithm creates fully grown trees top-down by splitting nodes. To split a node, a random

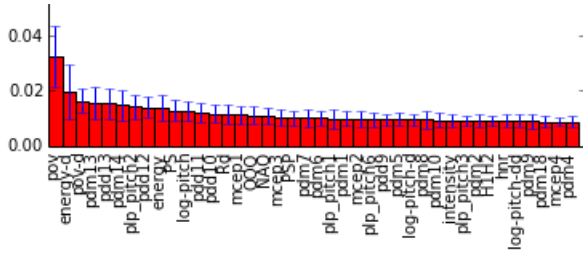

Figure 1: Feature salience for stød prediction (task 1). PLP are called *plp_pitch* and *energy* refers to C0 extracted with PLP features.

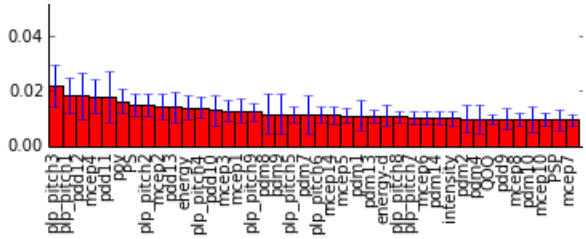

Figure 2: Feature salience for stød-bearing and stød-less phone variant discrimination (task 2).

subset $K = \sqrt{N}$ of all features N in the current node is selected as candidates for splitting criterion. For each k_j in K, a random cut-point a_j is chosen. The feature k_j with cut-point a_j which most improves entropy after a split is used to split the data in the node. Each decision tree is estimated on a random subset of the training data and we use sub-sampling with replacement to mitigate the under-representation of samples labelled with stød-bearing phones.

Relative feature importance can be ranked by the depth at which a feature is used to split a node because features used as splitting criterion closer to the root node contribute to the prediction of a larger fraction of samples. Final prediction is achieved by majority voting across all trees (1024). The samples are not weighted and classes are represented by an equal number of samples. This balancing is necessary to prevent the tree growing algorithm to favour features that predict a majority class.

5.1.1 Rankings

We rank features according to *salience* measured as mean reduction in entropy across the ensemble in two tasks: 1) binary stød prediction and 2) multi-class discrimination between stød-less and stød-bearing phone variants (e.g. [a$^?$] vs. [a], [m$^?$] vs. [m]) at sample level. Figures 1 and 2 show the 40 most salient features for each task.

A common set of features that are salient for phone discrimination and stød prediction emerges from studying Figures 1 and 2. The top 17 features are PLP 1-4, MFCC 1-4, PDD 10-13, PDM 13-14, PS, POV and log-pitch. In the following sections, SELECT will refer to this feature set and ALL will denote a set with all 120 features.

PDM and PDD are novel features in stød characterisation. That phase information is salient for stød prediction is to our knowledge a novel insight and interesting because PDM and PDD rank higher than many ASR-related features such as PLP-d, PLP-dd and some MFCC features. If this finding can be corroborated in the analysis of other corpora, phase features might be useful information to add to acoustic models in ASR.

6 Predicting stød from acoustic input

With acoustic features that are predictive of stød and reliable annotation, we can train classifiers that predict stød directly from acoustics with supervised training. The features in SELECT were also chosen for their ability to discriminate between stød-bearing and stød-less phone variants and in an ASR context, discriminating these phone variants will be sufficient to identify word class. Yoon et al. (2006) conducted a similar experiment for American English with *creak* to improve WER and they achieve an overall phone classification accuracy of 69.23% on 25 minimal phone pairs.

Following the same methodology, we train an SVM classifier with an RBF kernel and do not perform any optimisation of parameter values. We compare classifiers trained on ALL and SELECT to a baseline trained on PLP features in a balanced[8] 1v1 evaluation where the phone variants are only distinguished by the presence or absence of stød. We evaluate on JHP and do 5-fold cross validation on the training set because we cannot meaningfully separate a test set when we reduce the training data to 1/10th the original size.

We can see from Table 5 that the classification accuracy is much better than chance. The variance increases for all feature sets on JHP because it is much harder data, but all feature sets contain information that can help discriminate between stød-bearing and stød-less variants of the same phone, including standard PLP features.

[8] Balanced in terms of training data.

	Train	±	JHP	±
PLP	0.769	0.144	**0.713**	0.266
ALL	0.781	0.168	0.685	0.220
SELECT	**0.803**	0.176	0.600	0.104

Table 5: 5-fold cross validation on the training data across 40 phone variant pairs and mean classification accuracy on JHP across 5 pairs.

ID	Standard features	+stød	+pitch
1	PLP	×	×
2	PLP	✓	×
3	PLP	✓	✓
4	MFCC	×	×
5	MFCC	✓	×
6	MFCC	✓	✓

Table 6: The 6 conditions we test in this set of experiments.

7 Stød in ASR

We have discovered that acoustic features normally used in ASR contain information that signal the occurrence of stød and can then annotate stød in the pronunciation lexicon used in lexicon-based ASR systems to create a baseline with standard ASR feature input. We can then add stød-related features to ASR features to improve stød modelling and performance further. We split this set of experiments because adding more features to the training data could also have an adverse effect: we could be improving stød prediction at the expense of other speech sounds and because stød is relatively infrequent this could increase WER.

We trained several ASR systems with features where we augment MFCCs with features from SELECT and also tried to train AMs with SELECT and ALL as input. Training on ALL worsened performance and was a very expensive experiment, SELECT did not consistently perform better or worse and we have chosen to report on experiments where we observed performance improvements over more than one training run. The features log-pitch and POV from SELECT are good predictors of stød and are standard features to include in ASR for tonal languages together with Δ-pitch. These *pitch features* and modelling stød in the lexicon will be investigated in Section 7.1. Results with other features from SELECT that are not standard ASR features will be reported in Section 7.2.

We will no longer train on manually annotated data because we need more data than is available in DanPASS, PAROLE48 and JHP. We will train AMs and an LM on data from Språkbanken which is much larger and designed for ASR tasks.

7.1 Modelling stød in the lexicon

To train AMs, we use a pronunciation lexicon to convert text sequences to phone sequences. Phones are further subdivided to triphones and to state-tied HMM states or senones. The lexicon is central to state-of-the-art ASR and to test if stød can actually improve WER, we will use both a lexicon with stød annotation and without stød annotation. eSpeak generates phonetic transcriptions with stød by default and we simply remove the annotation in the first case.

We want to see if adding pitch features improve WER in ASR systems where stød is in the lexicon, so we test the six conditions in Table 6.

We base our recipe on the Wall Street Journal and Librispeech recipes in the Kaldi repository which trains a series of GMM models and a DNN model from scratch. We use IRSTLM (Federico et al., 2008) to train a language model (LM) on the training transcripts. We also tried to train a LM on ngram frequency lists calculated over 290 million words from Danish newspapers, but the performance degraded when we used the newspaper LM both on it's own and interpolated with the transcript LM and we conclude that the text genre is too different from our data sets. We use Matched Pairs Sentence-Segment Word Error (MAPSSWE) from the SCTK toolkit (Fiscus, 2007) to calculate statistical significance.

We train a GMM-based ASR system where we stack features in an ± 5 frame context and use LDA to project to 40 dimensions followed by a GMM with speaker-adaptive training using feature space MLLR (fMLLR) on top of LDA. The DNN is a 4-layer feed-forward network with 1024 nodes per layer and tanh-nonlinearities that we train on the same LDA+fMLLR transformed features. The learning rate starts at 0.01 and is decayed linearly to 0.001 over 15 epochs and trains for an additional 5 epochs at 0.001.

In Table 7, we see the impact on WER when we add stød annotation in the lexicon and add pitch features to the feature input (denoted +*pitch*) and evaluate on SPTEST. Adding stød consistently improves WER, but is not always statistically sig-

	PLP			MFCC		
AM	Baseline	+stød	+pitch	Baseline	+stød	+pitch
GMM+LDA	17.61	17.55	17.07‡	17.72	17.54	**16.88**‡
GMM+LDA+SAT	16.85	16.64*	16.49	17.14	16.81†	**16.17**‡
DNN	13.50	13.33	13.17	13.28	**13.08**†	13.38

Table 7: %WER performance on SPTEST. The best performance for each AM is in bold. Statistical significance over the condition in the column to the left is denoted by * if $p < 0.05$, † if $p < 0.01$ and ‡ if $p < 0.001$.

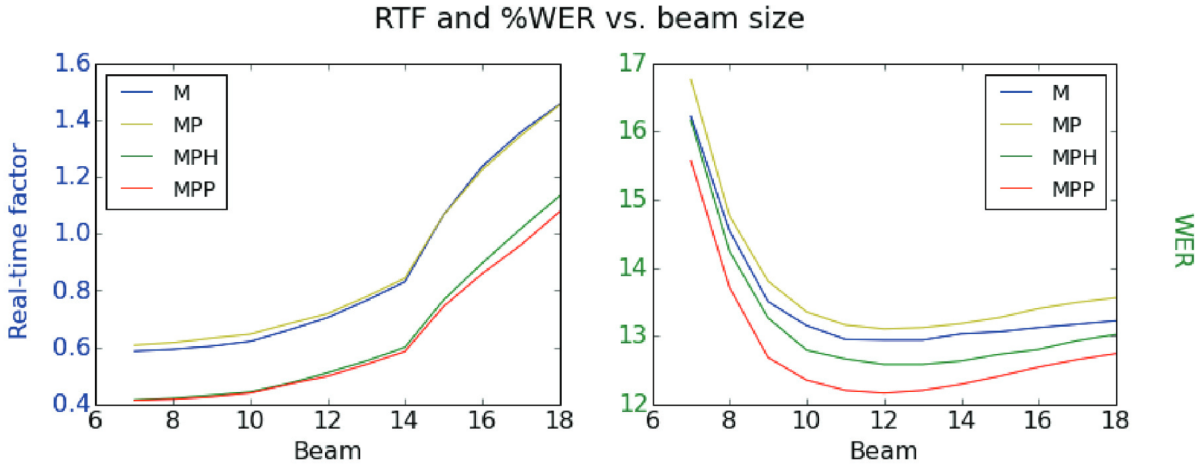

Figure 3: Beam parameters sweep on SPTEST. The optimal beam size is 12.

nificant. We found that adding stød annotation resolves many homophonic entries in the lexicon e.g. *hver, værd, vejr* and *vær* are transcribed as [vɛɐ] but stød resolves the ambiguity such that *værd, hver* and *vejr* transcribes as [vɛˀɐ] and *vær* as [vɛɐˀ]. Table 8 shows the impact of stød annotation on homophony in the pronunciation lexicon. The proportion of affected tokens in SPTEST, PAROLE48 and DanPASS are 27%, 26.7% and 7%, respectively and suggest that modelling stød can have a significant impact even though it appears infrequently.

Polygraphy	-stød	+stød	difference
4x	5	0	-5
3x	54	27	-27
2x	930	662	-268

Table 8: $2x$ denote the number of phonetic transcriptions that can be mapped to two words, $3x$ denotes the number of phonetic transcriptions that can be mapped to three words, etc.

Only when we evaluate the DNN AM trained in condition 6 do we not observe improved WER when we add pitch features. We also see that in general MFCC-based models outperform PLP-based models and that adding stød and pitch features gives a larger performance improvement in MFCC-based GMM AMs. The best performance is achieved with the DNN trained in condition 5.

We observe an interesting interaction between stød and pitch features in decoding speed. When we add stød and use the same decoder parameters, the real-time factor (RTF) becomes larger which means the ASR system takes more time to recognise a sentence. If we add pitch features to the acoustic input, decoding speed increases.

We encode stød-bearing phones as variants of stød-less phones when we estimate phonetic decision trees (PDT) and stød-bearing phones could become just an alias of the stød-less phone during state-tying because we do not increase the number of estimated probability density functions or leaves of the PDTs. We observe that state-tying tends to cluster together word position-dependent phones more often than stød variants such that clusters contain $[e_E^?, e_B^?]$, but not $[e_B]$[9]. There are 43-45 clusters of stød-bearing phones and 15-19 clusters of mixed stød-bearing and stød-less variants which indicate that stød often is a more important feature than word position.

[9]Subscripts denote word position

We can conclude that modelling stød explicitly in the pronunciation lexicon improves WER for both GMM and DNN AMs. There is a statistically significant overall WER improvement and we can conclude that there was no adverse effect on performance despite of the infrequent occurrence of stød. In all cases but one we also observe improvement from adding pitch features.

We find expert annotation is not necessary to take advantage of stød in ASR. We used a g2p-system to generate the pronunciation lexicon and can observe consistent performance improvements when we add stød annotation to the lexicon.

7.2 Stød-related acoustic features in ASR

We further investigate several features from SELECT: PDD 10-13, PDM 13-14 and Peak slope. Early experiments with GMM AMs showed significantly worse empirical results with Peak Slope and we chose to discard that feature from the rest of the experiments. We bin together PDD 10-13 and PDM 13-14 and denote them as *phase features*.

Abbr.	MFCC	Pitch	Extra
M	✓	×	×
MP	✓	✓	×
MPH	✓	✓	HRF
MPP	✓	✓	PDD10-13 PDM13-14

Table 9: Feature combinations and their abbreviations.

Harmonic Richness Factor (HRF) is a measure of harmonicity in the speech signal that Fernandez et al. (2014) use to improve ASR for Zulu and Lao and we expect a relevant measure to include in our study. We discard PLP features because AMs trained on MFCC features generally perform better than the PLP counterparts both in WER and RTF. We also include pitch features because they tend to improve performance and decoding speed and we need to estimate F0 to estimate both phase features and HRF.

The feature combinations we use are in Table 9 and we use early feature integration before LDA because it gave better performance in Metze et al. (2013) and worked well in previous experiments.

We will depart from standard test methodology and optimise one set of decoder parameters on SPTEST, DanPASS, and PAROLE48 which we also use as test sets. We could not find a method to completely isolate the impact the new acoustic features have on WER, but this is our best effort to reduce the impact from other factors. We randomly choose to optimise decoder parameters with the MP model. We will do a second evaluation where we sweep the decoder beam size and visualise the impact on RTF.

Table 10[10] shows that we can get better performance by adding HRF and phase features to the feature input. The improvement is significant on the multispeaker test sets, but not PAROLE48, where the MFCC baseline shows the best performance. The RTF constraint does not affect WER on SPTEST and Figure 3 shows that increasing the beam will have no effect, but we could increase the beam to 17 when we decode with MPH and MPP, but only to 14 with M or MP.

On PAROLE48, we see that M takes a small performance hit to maintain real-time decoding capabilities, but on DanPASS we can further improve MPP and MPH performance because the faster decoding speed allows us to use a larger decoding beam. For MPP, the improvement is significant at $p < 0.01$. The speed up in Figures 3 and 4 is constant for MPP and MPH compared to M. The MP RTF varies considerably and we cannot draw conclusions on the relationship to MPP and MPH based on these experiments.

We can conclude that HRF, PDD10-13 and PDM 13-14 are beneficial acoustic features to use in Danish ASR. WER decreases and decoding speed increases when we add these features. While phase features seem to provide the best improvements, the phase feature extraction method is slower than real-time and our current recommendation is to use HRF. Notes in the Covarep source code suggest that real-time phase feature extraction is possible at the cost of precision, but implementing real-time phase feature extraction is beyond the scope of this research.

8 Conclusion

We discovered that stød annotation is reliable when it is annotated by expert phoneticians and used this insight to discover predictive acoustic features that are novel in the phonetic characterisation of stød.

We also discovered that we do not need expert annotation to use stød in ASR to improve

[10] See Table 1 for summaries of the corpora used as test sets.

Features	SPTEST	PAROLE48	DanPASS
M	12.94	**29.78** **(29.89)**	53.83
MP	13.10	30.38	54.73
MPH	12.58‡	30.38	51.06 (50.46)
MPP	**12.16**‡	30.05	**49.02*** **(48.79*)**

Table 10: WER performance using the same decoder parameters for each test set. WER under the $RTF < 1$ constraint are in parentheses if different. Statistical significance is compared to M. These numbers are not directly comparable to Table 7.

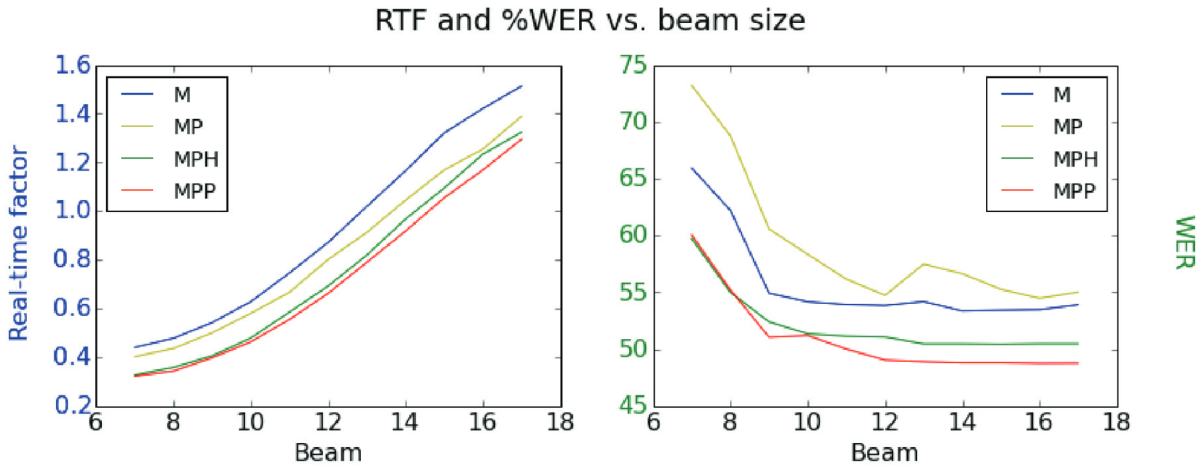

Figure 4: Beam parameter sweep on DanPASS.

performance. The harmonic richness factor and the phase features PDD10-13, PDM13-14 also improve ASR performance and this indicates we have successfully modelled stød explicitly in the lexicon and implicitly with predictive acoustic features without degrading overall performance. We believe that these features can improve performance in absence of stød annotation.

We tried to predict stød as a binary classification task, i.e. predict the presence or absence of stød regardless of the co-occurring phone, but this was not possible because creaky voice, laryngealisation and other acoustic signals that correlate with stød also occur when there is no stød-bearing phone. In future work, we want to experiment with pronunciation variants based on stød to accurately model the optional nature of stød and do an ablation study where we use more features from SELECT, and do not include pitch features in the feature input. We also need to investigate what impact these features have in the absence of stød annotation.

We used open source software and features from ASR and speech analytics so our experiments can be reproduced and reapplied to Swedish and Norwegian. Språkbanken also includes Swedish and Norwegian and eSpeak can generate pronunciations with tones for both languages.

There are no previously published results on Språkbanken or any of the test sets and this was state-of-the-art performance in early 2016. New state-of-the-art performance on Språkbanken-test[11] also model stød in the lexicon.[12]

Acknowledgments

This work was supported by the Danish Agency for Science and Higher Education, Copenhagen Business School and Mirsk Digital ApS. We thank Klaus Akselsen, Peter Juul Henrichsen, Dirk Hovy and Srinivas Bangalore for support and mentoring through the long process.

[11] See https://github.com/kaldi-asr/kaldi/blob/master/egs/sprakbanken/s5/RESULTS

[12] See https://github.com/kaldi-asr/kaldi/blob/master/egs/sprakbanken/s5/local/dictsrc/complexphones.txt

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluation phonemic transcription of low-resource tonal languages for language documentation. In *LREC*.

P. Boersma. 2002. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 960–964. IEEE.

Jonathan Duddington. 2010. *eSpeak Text to Speech*. Web publication: http://espeak.sourceforge.net/.

Gunnar Fant. 1995. The LF-model revisited. Transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, 2(3):40.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Interspeech*, pages 1618–1621.

Raul Fernandez, Jia Cui, Andrew Rosenberg, Bhuvana Ramabhadran, and Xiaodong Cui. 2014. Exploiting Vocal-Source Features to Improve ASR Accuracy for Low-Resource Languages. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Eli Fischer-Jørgensen. 1989. *A phonetic study of the stød in Standard Danish*. University of Turku, Phonetics.

J Fiscus. 2007. Speech recognition scoring toolkit ver. 2.3 (sctk).

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning*, 63(1):3–42.

Nina Grønnum. 2005. Fonetik og Fonologi, 3. udg. *Akademisk Forlag, København*.

Nina Grønnum. 2006. DanPASS-a Danish Phonetically Annotated Spontaneous Speech corpus. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genova, Italy, May*.

Nina Grønnum, Miguel Vazquez-Larruscaín, and Hans Basbøll. 2013. Danish Stød: Laryngealization or Tone. *Phonetica*, 70(1-2):66–92.

Gert Foget Hansen. 2015. *Stød og stemmekvalitet: En akustisk-fonetisk undersøgelse af ændringer i stemmekvaliteten i forbindelse med stød*. Ph.D. thesis, Københavns Universitet, Faculty of Humanities, Department of Nordic Research. In Danish.

Peter Juel Henrichsen. 2007. The Danish PAROLE corpus-a merge of speech and writing. *Current Trends in Research on Spoken Language in the Nordic Countries*, 2:84–93.

Peter Juel Henrichsen and Thomas Ulrich Christiansen. 2012. Speech Transduction Based on Linguistic Content. In *Joint Baltic-Nordic Acoustics Meeting, Odense, Denmark*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard H Hovy. 2013. Learning Whom to Trust with MACE. In *HLT-NAACL*, pages 1120–1130.

Andreas Søeborg Kirkedal. 2013. Analysis of phonetic transcriptions for Danish automatic speech recognition. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013) May 22–24, 2013, Oslo University, Norway.*, NEALT Proceedings Series 16.

Andreas Søeborg Kirkedal. 2014. Automatic Phonetic Transcription for Danish Speech Recognition. *CRITT-WCRE Conference*.

Aditi Lahiri, Allison Wetterlin, and Elisabet Jönsson-Steiner. 2005. Lexical specification of tone in north germanic. *Nordic journal of linguistics*, 28(1):61–96.

Shang wen Li, Yow-Bang Wang, Liang-Che Sun, and Lin-Shan Lee. 2011. Improved tonal language speech recognition by integrating spectro-temporal evidence and pitch information with properly chosen tonal acoustic units. In *INTERSPEECH*.

Florian Metze, Zaid A. W. Sheikh, Alexander H. Waibel, Jonas Gehring, Kevin Kilgour, Quoc Bao Nguyen, and Van Huy Nguyen. 2013. Models of tone for tonal and non-tonal languages. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 261–266.

Taniya Mishra, Andrej Ljolje, and Mazin Gilbert. 2011. Predicting human perceived accuracy of ASR systems. In *Twelfth Annual Conference of the International Speech Communication Association*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Tomas Riad. 2000. The origin of danish stød. *Analogy, Levelling and Markedness: Principles of Change in Phonology and Morphology*, pages 261–300.

Niklas Vanhainen and Giampiero Salvi. 2014. Free Acoustic and Language Models for Large Vocabulary Continuous Speech Recognition in Swedish. *training*, 965(307568):420–8.

Tae-Jin Yoon, Xiaodan Zhuang, Jennifer Cole, and Mark Hasegawa-Johnson. 2006. Voice quality dependent speech recognition. In *International Symposium on Linguistic Patterns in Spontaneous Speech*. Citeseer.