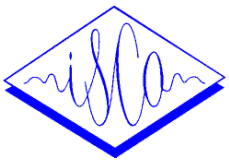

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

The multifarious r - sound.

Knut Kvale

Div. of Electrical Engineering
and Computer Science
The Norwegian Institute of Technology
N-7034 Trondheim, NORWAY
Telefax: +47 7 59 26 40
e-mail: kvale@tele.unit.no

Arne Kjell Foldvik

Dept. of Linguistics
University of Trondheim
N-7055 Dragvoll
NORWAY
Telefax: +47 7 59 61 19
e-mail: arne.foldvik@avh.unit.no

10.21437/ICSLP.1992-339

ABSTRACT

The most common pronunciation of /r/ in Norwegian is as an apical alveolar tap, i.e. the tongue tip touches the alveolar ridge and makes a short closure phase. In a manual segmentation task [1] we thus assumed that the tongue movement for the alveolar tap is a symmetric one. In our segmentation of /r/ we therefore included 10ms on each side in addition to the segment of less intensity in the spectrogram to include the tongue tip movement. In this paper we will demonstrate that different phonemic contexts systematically affect the realisation of /r/ and how this leads to some exceptions from the 10ms-addition-rule mentioned above.

Depending on the speaker's dialect background, the /r/-phoneme in Norwegian is produced as an apical tap or trill, a uvular tap or trill, or a post-palatal, velar or uvular fricative [2]. Recent studies [2] show a widespread and rapid change from an apical to a dorsal pronunciation of /r/ in South/West Norwegian dialects. Reasons for this ongoing change will be discussed.

Based on the annotated European multilingual EUROM.0 speech database we will discuss /r/ pronunciations in different languages and how /r/ has been segmented.

1. INTRODUCTION

For basic speech research and for the development and assessment of automatic speech recognition and text-to-speech systems carefully designed and annotated multilingual speech databases containing samples of different speakers and speaking styles are needed. Huge databases give more reliable statistical data, generalisations can be drawn, and general acoustic-phonetic speech knowledge is more easily obtained. Thus, within the European ESPRIT-SAM project [3] the EUROM.0 and EUROM.1 speech databases and in the American DARPA project [4] the TIMIT speech database have been compiled.

The problem is how to segment and label this speech material in order to make it useful for the tasks mentioned above. Segmentation is the process of dividing the continuous speech pressure waveform into discrete, non-overlapping, and directly succeeding discrete entities, often called traditional or linear segmentation. By labelling we mean a description of a given segment as defined above. Annotation is here used as a cover term for segmentation and labelling.

Neither the articulatory processes nor the acoustic speech signal are composed of discrete segments. Adjacent gestures overlap and merge and so do the phonemic cues in the corresponding speech signal. In this sense segmentation of speech is impossible. However, since annotated speech databases are needed in speech technology and research, (phonemic) segmentation is carried out even if theoretical doubtful compromises have to be reached in order to do so.

For the EUROM.0, native phonetician(s) for each language segmented and labelled the speech corpus in terms of the computer readable SAM Phonetic Alphabet (SAMPA) symbols [5]. SAMPA is phonemic and thus used according to the analysis of distinctive sound oppositions within each language. The acoustic realisations covered by one SAMPA symbol will vary between languages and within a language it will cover all allophones of that phoneme. However, some symbols can be used to represent allophonic variants within a language, e.g. apical and dorsal realisation of /r/ can be transcribed with r and R respectively. The SAMPA symbols were selected because the SAM project aimed at developing a standard (semi-)automatic speech annotation strategy [3] and intended to use the annotation for further analytic work [6].

In the manual segmentation and labelling of the Norwegian EUROM.0 recordings our convention was that if we heard a sound in context it should be segmented and labelled, even if we could not see any acoustic cues in the waveform or the spectrogram or could not hear the sound in isolation. The audible, but invisible phoneme was then squeezed in between the two surrounding phonemes by taking a couple of pitch periods from each of them (for further details, see [1]).

In the process of manual segmentation and labelling of speech we investigated the criteria that were applied when placing boundaries and noticed which were rather reliable and which were not. This is obviously very important and valuable knowledge when training and evaluating systems for automatic segmentation of speech and for the development of standard multi-lingual annotation procedures.

Due to the great variability and intrinsic complexity of the speech signal it is difficult to make general guidelines for the annotation of speech. In this paper we exemplify this by concentrating on the r-sound. The /r/ is realised differently in different languages and dialects and often has allophonic contextual variants within one and the same language. Therefore, conventions have to be made to include these variants.

2. HOW CONTEXT AFFECTS THE REALISATION OF THE NORWEGIAN r-SOUND

The most common pronunciation of /r/ in Norwegian is as an apical alveolar tap, i.e. the tongue tip touches the alveolar ridge and makes a short closure phase. On the spectrogram the closure phase of the tap shows up as a low frequency voicebar very similar to the closure phase of the retroflex, lateral flap /rL/. Figure 1 below depicts the waveform and the broad band spectrogram of the words "åre" /O:re/ (=oar) and "åle" /O:rLe/ (=crawl). In a manual segmentation and labelling task [1] we thus assumed that the tongue movement for the alveolar tap is a symmetric one, i.e. the apex takes the same time to move up to the alveolar ridge as away from it after the closure phase. We therefore included 10 ms, or approximately one pitch period, on each side of the /r/-segment in addition to the portion of less intensity in the spectrogram to include the tongue tip movement.

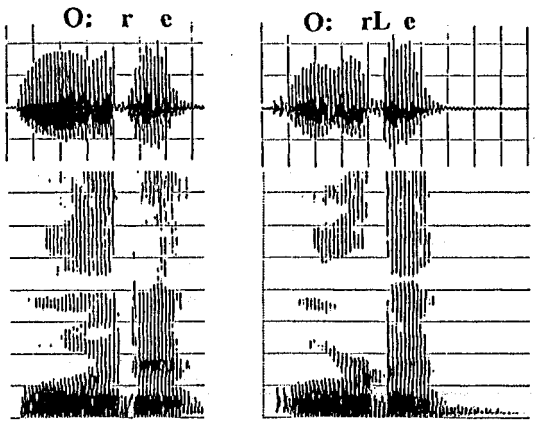

Figure 1 *Waveform and broad band spectrogram of /O:re/ and /O:rLe/ pronounced in isolation by a male speaker. Neighbouring vertical lines are 50ms apart. In the spectrogram the frequency difference between neighbouring horizontal lines is 1 kHz.*

However, the symmetry was only found for /r/ in intervocalic position and in vowel-/r/-voiced fricative /v/ position. Accordingly, some few exceptions from the 10ms-addition-rule mentioned above had to be made:

A. After a plosive burst or a voiceless fricative as in the words "tro" /tru:/ (=belief), "dro" /dru:/ (=left) and "fri" /fri:/ (=free) a period of voicing and formant structure is seen in the spectrogram before the tongue tip reaches the alveolar ridge. This is also the case for /r/ in sentence initial position as in "ren" /re:n/ (=clean). This voiced portion which was often much longer than the 10 ms we defined for the symmetrical /r/ realisation was included in the /r/ segment, as seen in figure 2. In some words, especially where a voiceless plosive precedes the /r/, as in "prinsipp" (=principle) and "prate" (=chat) the voiced period is perceived as an epenthetic @ sound between the plosive and /r/. This @ was not transcribed but included in the /r/ segment.
Segmenting the speech signal into twice as many acoustic subwords as phonemes with our automatic segmentation program [7], the @ and the closure portion of /r/ is segmented into two segments.
For some speakers the /r/ may also be partially devoiced particularly if the preceding phoneme is an unvoiced sound.

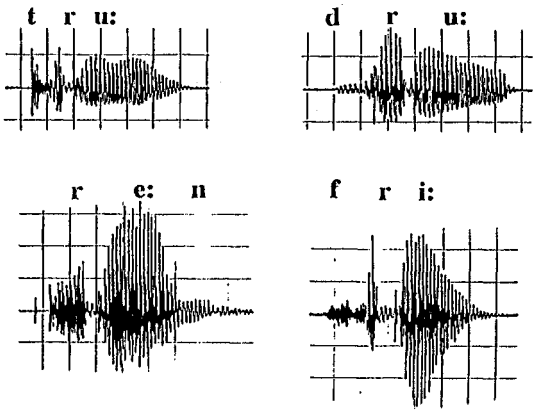

Figure 2 *Waveforms of /tru:/ and /dru:/, /re:n/ and /fri:/ pronounced in isolation by a male speaker. Neighbouring vertical lines are 50ms apart.*

B. When /r/ precedes a voiceless fricative, voiceless plosive or a nasal, a short period of voicing appears after the /r/ closure as seen in figure 3. This effect may also occur at the end of sentences.

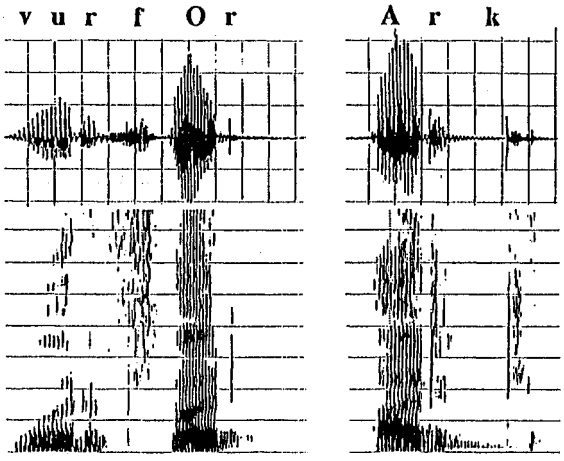

Figure 3 *Waveform and broad band spectrogram of "hvorfor" /vurfOr/ (=why) and "ark" /Ark/ (=sheat) pronounced in isolation by a male speaker. Neighbouring vertical lines are 50ms apart. In the spectrogram the frequency difference between neighbouring horizontal lines is 1 kHz.*

In figure 2 and 3 we notice that even when one side of the /r/ does not fulfil the symmetric assumption one pitch period of the neighbouring phoneme is still included in the /r/-segment.

C. At the end of sentences the amplitude of /r/ normally decreases evenly towards the end as seen in /vurfOr/ in figure 3, and -/er/ in figure 4. If the voice source is turned off early the last part of /r/ becomes devoiced. This voiceless part was also included in the /r/-segment.

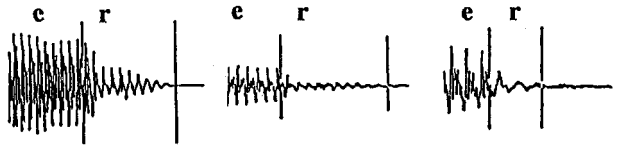

Figure 4 *Waveform of -/er/ pronounced by three speakers. Taken from the end of a longer word at the end of an utterance. Time axis differs from that in the other figures.*

3. REGIONAL DIFFERENCES IN r-PRONUNCIATION IN NORWEGIAN

There is no standard pronunciation of Norwegian and the pupils' right to use their own dialect pronunciation in school is stated in education laws. Depending on the speaker's dialect background, the /r/-phoneme is produced as an apical tap or trill, a postalveolar approximant, a uvular tap or trill, or a post-palatal, velar or uvular fricative or approximant [2].

In figure 5, the apical alveolar tap and the dorso uvular approximant are compared intervocalically. The uvular approximant with no clear acoustic boundary cues makes the segmentation more difficult. The uvular /r/ when preceded or followed by an unvoiced sound is realised as a fricative.

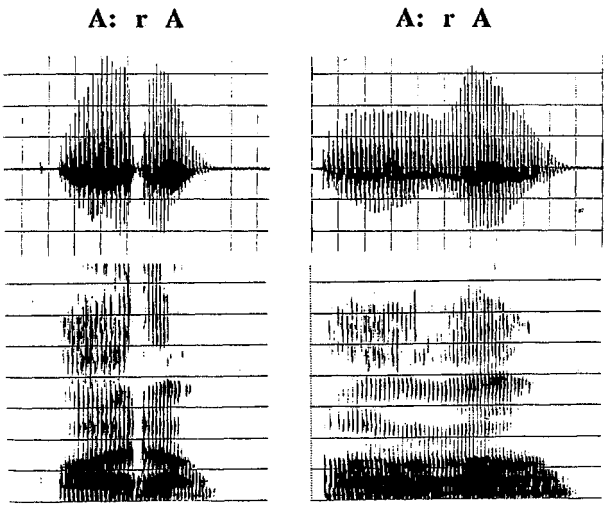

Figure 5 *Waveform and broad band spectrogram of /A:rA/ pronounced in isolation by a male speaker, with an apical alveolar tap to the left and an uvular dorsal approximant to the right, which can be considered as the two articulatory extremes of Norwegian /r/-pronunciation.*

During this century the change from apical to dorsal pronunciation of /r/ has affected and is still affecting areas particularly in Southern and South-Western Norway [2]. Dorsal /r/ has spread to areas where as many as $^1/_8$ of the population of the country live, from dorsal-r towns into the surrounding rural districts.
Although the dorsal /r/ motorically is easier to pronounce than the apical /r/ this can not account for the spreading alone. The dorsal /r/ spreading is facilitated by dorsal and apical /r/ being equally socially acceptable, and also by the fact that on the whole people do not experience a change in /r/ pronunciation as a change of dialect.
The main reason, though, for the dorsal /r/ spreading is supposed to be the prestige connected with dorsal /r/-towns and the linguistic influence that these centres exert on the rural districts.
There is reason to believe that the dorsal /r/ pronunciation will spread only as far as this influence from dorsal towns goes. And there are no signs of dorsal /r/ spreading to any of the apical /r/ towns.

4. COMPARISON WITH OTHER LANGUAGES

Many manual annotation strategies have been proposed, e.g. for EUROM.0: Norwegian [1], Danish [8], Italian [9], English [10], Swedish [11], French [12] and TIMIT: American English dialects [13]. Since no general guidelines exist for manual segmentation and labelling, everyone makes his or her own conventions. These conventions are based on what is felt natural according to the phonetic school of the labeller and what is the intended use of the annotated speech material. If the conventions are strictly adhered to, the segmentation and labelling is correct.

For EUROM.0 the Swedish, Norwegian and English annotations are performed at a fairly similar level of labelling, defined as the broad phonetic level within SAM [14]. Each phoneme in the perceived phoneme sequence is segmented whether acoustic cues are seen in the waveform or not. However, if acoustic cues are seen the boundaries are placed at these. The annotation of the Danish EUROM.0 and the TIMIT database is performed at some acoustic-phonetic level where an abrupt change in the signal is marked whether it indicates a phoneme boundary or not. In manual cross-comparison tests or in the assessment of automatic speech segmentation algorithms these differences in strategy have to be taken into account. The acoustic-phonetic approach will presumably give much better correspondence to automatically placed boundaries than the phonemic approach.

Returning to the /r/ the Swedish database shows considerable variation in the /r/ pronunciation. It is claimed [11] that /r/ in vowel context is one of the most impossible to segment because its cues are superimposed on the vowels. We found at least three different /r/-realisations in Swedish for r intervocalically: as an apical approximant with or without abrupt change in amplitude when going from and to the surrounding vowels, or as a partly devoiced one.

When an unvoiced plosive is followed by an /r/ the Swedish database indicates that a partly devoiced tap is the most frequent /r/ pronunciation. In the segmentation parts of the voiceless burst phase have been included in the /r/ segment. An epenthetic @ may also occur before the /r/. This is included in the /r/-segment. In both cases this segmentation approach is similar to the Norwegian one.

The Italian /r/ is realised as an apical tap and the geminate /rr/ as a trill. The pronunciation is similar to Norwegian /r/ pronunciation intervocalically and when preceded by a fricative and voiceless plosive. But /r/ in sentence initial position differs from Norwegian in the same context. The Italian segmentation of /r/ is on the whole consistent with the Norwegian approach and the /r/ realisations for the two languages can thus be combined to enlarge the language specific training material for our statistically based automatic segmentation method [7].

The most common /r/ realisation in British English is a (post) alveolar approximant, with glide-like features [10]. This /r/ is segmented at the "half-way point in the F_3 glide between its maximum and minimum values" [10].
In the English EUROM.0 recordings the /r/ in vowel context, especially after non-open front vowels as in "be written" /birItn/, is realised as an approximant with less amplitude seen in the waveform. However, since the half-way convention is applied, more of the neighbouring vowels is included in the /r/-segment than would be the case if the Norwegian approach had been adopted.

/r/ in Danish, German, English and French is realised differently from /r/ in the Norwegian database which only exemplifies apical /r/ pronunciation. Norwegian dorsal /r/ dialect speakers will show /r/ pronunciations which are similar to /r/ in German and French.

In a cross-comparison test [6] a Dutch phonetician aligned a given phonemic transcription with speech by indicating the centres for each phoneme for some sentences of the English, French and Danish EUROM.0 speech material. (The labeller spoke perfect English, moderate French but had no knowledge of Danish). The centre positions were compared to the ones placed by native phoneticians in each language.
The deviations found were explained as a result of different criteria in each language, random variation in the criteria, and the use of different equipment. /r/ was reported to cause the labellers comparatively many difficulties.
The problems that the labeller encountered may have been caused by different acoustic manifestations across the languages. In addition, segment boundaries had been placed differently for different languages even if the /r/-realisation turned out to be fairly similar. Another reason for the difficulties with defining the centre of this phoneme may have been that many /r/-allophones are not symmetrical and give different results depending on which acoustic parameters the decision is based on.
The experiment shows that segmentation of speech is particularly difficult when the labeller does not know the common realisations of the phonemes in the actual language.

5. SUMMARY

Contextual affects on the Norwegian apical alveolar tap /r/-realisation are discussed and conventions for segmenting it are described. Regional differences are shown and a widespread and rapid change from apical to dorsal pronunciation of /r/ in South/West Norwegian dialects is discussed.

We have investigated allophonic variation in /r/ pronunciation across languages and looked at differing segmentation procedures. Since manually segmented speech databases are used as basis for evaluation of automatic segmentation algorithms, it is important that shortcomings and inconsistencies in manual segmentation is highlighted.

We would like to argue that criteria for standardised segmentation approaches have to be established for future multi-lingual speech databases.

6. REFERENCES

[1] K. Kvale and A.K. Foldvik, "Manual Segmentation and Labelling of Continuous Speech", Proc. ESCA Workshop on "Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication", pp. 37.1-5, Barcelona, 1991.

[2] A.K. Foldvik, "The change from apical to dorsal r in Norwegian". Proc. 11th International Congress of Phonetic Sciences, Tallinn, Vol.1, pp. 177-178, 1987.

[3] A. Fourcin, "Linguistic engineering and speech assessment methods" in *Speech, hearing and language, work in progress 1991*, vol 5, University College London, Department of Phonetics and Linguistics, pp. 65-74, 1991.

[4] S.P. Pallett, "Speech corpora and performance assessment in the DARPA SLS program", Proc. International Conference on Spoken Language Processing, Kobe, Japan, pp. 24.3.1-4, 1990.

[5] J.C. Wells, W. Barry, M. Grice, A. Fourcin, D. Gibbon, "Standard Computer-Compatible Transcription", SAM stage report sen. 3, SAM-UCL-037, in *ESPRIT PROJECT 2589 (SAM): Final Report; Year Three; 1.3.91-28.2.92*, 1992.

[6] A. Van Erp, M. Grice and W. Barry, "Manual Labelling of Danish, Dutch, English and French Speech Material on EUROM.0", in *SAM, EXTENSION PHASE, FINAL REPORT, 1 April 1988 - 28 February 1989.*

[7] T. Svendsen and K. Kvale, "Automatic alignment of phonemic labels with continuous speech". Proc. International Conference on Spoken Language Processing, pp. 997-1000, Kobe, Japan, Nov. 1990.

[8] N. Dyhr, O. Andersen and P. Dalsgaard, "Labelling criteria for the Danish EUROM.0 database", SAM-document no.: SAM-IES-055, 1992.

[9] P. Cosi, "Segmentation and Labelling of EUROM.0 Italian Continuous Passage", Internal Report for ETR group of SAM, Document No. SAM-CSRF(CNR)-04.

[10] W.J. Barry, "Labelling criteria: phonemic and acoustic-segment labelling", Document no. SAM-UCL-026, for workgroup ETR, oct. 1990.

[11] L. Nord, "Report on manual labelling criteria used on the Swedish EUROM.0 material", ETR-document, Sept. 1990.

[12] D. Autesserre, G. Perennou and M. Rossi, "Methodology for the Transcription and Labelling of Speech Corpus", Journal of the International Phonetic Association, Vol. 19, no. 1, pp. 2-15, July, 1989.

[13] S. Seneff and V.W. Zue, "Transcription and Alignment of the TIMIT Database", in *Getting Started With The DARPA TIMIT CD-ROM*, chap.4. Unpublished manuscript ("To be distributed with the TIMIT database by NBS"), 1988.

[14] W.J. Barry and A.J. Fourcin, "Levels of labelling" in *Speech, hearing and language, work in progress 1990*, vol 4, University College London, Department of Phonetics and Linguistics, pp. 31-43, 1990.