DEVELOPMENT OF RULES FOR AUTOMATIC RECOGNITION OF NASAL CONSONANTS

S Nulsen, D Landy, M O'Kane, P Kenne and S Atkins

Faculty of Information Sciences and Engineering
University of Canberra

ABSTRACT - Examination of speech waveforms has shown that the nasal consonants are one of the easiest classes of sounds to recognise by eye. What is it that so clearly distinguishes these from the other sounds? Three features may be identified:

- their characteristic shape,
- an amplitude which is much lower than nearby vowels but which is well above zero and cannot be confused with noise,
- a very clean smooth waveform, frequently almost sinusoidal. This is indicative of the concentration of spectral energy in the low frequency region, which has previously been well documented.
- a set of recognition rules for the class of nasal consonants was developed to capture these observations using the WAL speech recognition programming environment. The development of these rules is discussed and results are presented for three speakers.

INTRODUCTION

The nasal recognition rules were developed using the WAL speech recognition programming environment. The WAL speech recognition programming environment (see paper by Kenne, Landy, O'Kane, Nulsen, Mitchell & Atkins in this volume) uses speech recorded at 20kHz. Functions of the speech wave referred to as **derivatives**.

An envelope derivative is obtained by taking the maximum value of the speech wave occurring in a fixed window of time. For this to give the desired result the window must exceed the fundamental frequency of the speech. Subsequent points of the derivative are found by the same method at intervals of half the window. This derivative is known as **envtop**. A similar derivative, **envbot**, is found by taking the minimum values. Since the speech wave is not symmetric this is not simply the negative of envtop. The similarities between the upper and lower envelope derivatives can be picked out using a further derivative, **abssum**, which is the sum of the absolute values of envtop and envbot.

The slopes of the upper and lower envelope derivatives are given by the derivatives **envtslope** and **envbslope**. These are the rise or fall between two successive envelope values divided by the time interval between them.

Frequency information can be obtained from a fast fourier transform among other derivatives. A derivative called **fftband** sums the amplitudes of the fourier transform points in a specified frequency range at some time and gives this as a percentage of the sum of all the fourier amplitude points at that time.

A further derivative **fftbandlog** is calculated in a similar way but uses the logarithm of the fourier transform amplitudes above a certain cut-off amplitude. This derivative is useful for distinguishing between frequency regions of low amplitude.

THE AUTOMATIC SPEECH RECOGNITION RULES

WAL's automatic speech recognition rules detect when the values of a derivative fall above or below specified values or in a specified range of values. When this happens the rule is said to **fire**. Each continuous period of time over which a rule fires is referred to as a **segment**. A rule may be modified to select only segments of a specified time range. Segments may also be extended by a specified length. This can be used to connect a series of short segments. More complex rules may be obtained by combining simpler ones using the connectives **and**, **or**, **then**, **before**, **after** and **not**.

RULE DEVELOPMENT TECHNIQUE

The WAL speech recognition system relies on the ability of the rule developer to pick out patterns in the speech wave and its derivatives. Once these have been found they can easily be converted into recognition rules.

Several sentences containing a large number of nasal consonants in various contexts were devised. The two sentences used here are:

> Sentence n1 (20 nasal consonants)
> N1: While in the snow calm Noreen Manson saw a small animal gnawing an amber Malawian elm.
>
> Sentence n2 (18 nasal consonants)
> N2: The minimum angle of the handle gave ample room for dismantling Norm's family sewing machine.

These sentences also contain examples of other phonemes previously found to cause rules attempting to recognize nasals, to fire falsely. These phonemes include the liquid /l/, the glide /w/ and the vowels /ɔ/ (as in port) and /u/ (as in boot). Both sentences are about 6 or 7 seconds long.

One wave was taken and examined closely. This was an example of the first sentence spoken by a female speaker. Various sets of rules to try to recognize the nasal consonants were written and run over this wave. These rules were modified or abandoned until a set of rules with a high detection rate and a low false firing rate was obtained. This set of rules was then tested over a second sample of speech by the original speaker, and over the same sentences spoken by a second female speaker and by a male speaker. In all fourteen sets of rules were tried over the original wave before reasonable results were obtained.

THE SHAPE OF THE NASAL CONSONANTS

The shape of the nasal consonants depends on the broad phonetic context in which they occur. On this basis they may be divided roughly into four main groups.

The first group will be referred to as **initial** nasals. See Figure 1. These occur not just at the beginning of a syllable but where the nasal also follows a period of very low amplitude such as may occur during a pause in the speech after a plosive, or in a fricative like /f/. These nasals may be of quite short duration. The envelope of the nasal segment exhibits a convex shape, rising steeply from close to zero to flatten at some low amplitude in a short period of time. This is usually followed by a very sudden increase in amplitude as the following vowel begins.

The second group, the **final** nasals, may be regarded as the time reversal of the initial nasals. See Figure 2. Following a sudden drop from the preceding vowel, the envelope of the nasal segment also takes on a convex shape. Over a short period of

time it curves slowly from almost horizontal to fall steeply to zero. This is then followed by some period of zero or very low amplitude.

The third group is the **medial** nasals. See Figure 3. These typically occur between two vowels. At each end of the nasal segment the amplitude of the speech waveform falls or rises rapidly to the adjacent vowel but in between the envelope is held steady, close to a constant value for some time. This time interval is usually longer than that of the initial or final nasal segments. This contrasts with the dips formed by, for example, the glide /w/ or the voiced fricative /v/ where the slope of the envelope is always changing, with a concave shape, and no constant value is maintained.

The fourth main group is the **syllabic** nasals. These are like a initial nasal followed immediately by a final nasal. They generally do not have any flat section but form a smooth convex hump. See Figure 4.

Initial and final nasals may be further subdivided, **long final** nasals forming another group. See Figure 5. These occur in a similar context to the final nasals but where the nasal segment is more stressed. They can be considered as a combination of medial nasal followed by a final nasal. The nasal segment consists of a period of time over which the envelope of the waveform is flat, with zero slope, or has a very small constant negative slope, followed by a short period of rapidly increasingly negative slope as the amplitude drops to zero. The lower envelope of the waveform may not have an exactly corresponding shape to the upper. Where the upper has a small negative slope the lower may be perfectly flat. The division between the nasals and the long nasals is clearly not well-defined but a long final nasal may be said to occur whenever a final nasal begins with a clear period of zero or very small constant slope of its envelope.

Similarly **long initial** nasals may be considered a combination of initial nasal followed by medial nasal. See Figure 6. There is generally less variation in the duration of initial nasals. Very long initial nasals do not occur frequently. However initial nasals tend to be stressed more, so very short initial nasals are also less common. Most initial nasals will fall into this group of long initial nasals.

Trailing nasals are yet another subdivision of the final nasals. See Figure 7. These are unstressed final nasals occurring before a long pause. They have a low amplitude with the "average" amplitude falling away gradually over a long period, but may be interrupted by brief periods of higher amplitude.

RULE DEVELOPMENT

Rules using the envtslope and envbslope derivatives were written to detect regions of flatness in the speech wave by specifying that the slope of the envelope should fall within a specified range about zero for a certain length of time (eg 20ms). Other rules were written to detect slopes above or below certain values for periods of time between 10 and 40ms. These rules wre then combined in various orders using the then, or before and after connectives in attempts to recognize the shapes described previously. This had only limited success.

A large part of the failure was due to inconsistancy of the envelope derivatives as they are currently defined. Results can vary significantly depending on the positioning and length of the windows used in the calculation of the derivatives. Artificial flat spots occur in the enveloope if the same point provides the extreme value in two consecutive windows. This is not uncommon since the windows overlap by half their length. Generally this is not a problem when the envelope derivatives are used directly, but when the slope is calculated from them the flat spots are translated into sudden dips to zero surrounded by values of enlarged magnitude!

In order to overcome the problems involved in using the envelope slope derivatives a series of rules was written based directly on the abssum derivative. (This derivative is the sum of the magnitudes of the upper and lower envelopes.) Each rule in the series specified that abssum must lie within a small range for at least 16ms. There were twelve rules in the series, the range of each rule overlapping that of its neighbours. When all the twelve were combined using the or connective, the resulting rule detected regions in which the envelope of the speech wave could only vary slowly and its magnitude fell in the range of values observed for the nasals. This rule missed a short initial and a short final nasal as well as a final trailing nasal. It also picked up quite a large amount of other speech.

To reduce the number of false detections the number of rules in the series was increased to 23 and the range of values specified in each rule was halved. The times for which abssum must remain within each range was increased to 20ms for ranges of lower amplitudes and to longer times for the ranges of larger amplitudes. The more stressed nasals are generally both longer and higher, providing a correlation between duration and magnitude.

The resultant more stringent rule did indeed eliminate a large amount of the extraneous material, but it still detected a significant amount extra. It also failed to recognize another short initial nasal, and a very short medial nasal, meaning that altogether it missed five of the twenty nasals in the speech sample used.

No set of rules using shape alone will be sufficient to recognize the nasal consonants. The liquid /l/ in particular often looks very similar, especially in outline. Frequency information must also be taken into account. A very simple frequency rule, which requires that most (75%) of the fft amplitude calculated be in the range 50 to 400Hz, has been found to be effective in detecting the nasals without picking up /l/'s.

When this frequency rule was combined with the previous amplitude-based rules using the and connective, false detections were still made in several places. These include the glide between the final vowel sounds in the word "Malawian". This was eliminated by including a requirement that the contribution from the frequency range 900 to 1600Hz be very low. Because this is also small for the nasals it was necessary to use the fftlogband derivative.

When the rules were tried on a second passage of speech another false detection was made where an /l/ ran into an /r/. Adding another fftband rule requiring the contribution of frequencies above 500Hz to be less than 10% eliminated this.
It was noted that the resulting frequency rule alone proved so effective that the only false detections in the sample of speech under examination were occurring in regions of very low amplitude speech. So the long and complicated series of amplitude rules using abssum were put aside and replaced by a simple amplitude rule requiring abssum to be greater than some small value.

The biggest remaining problem was that the results from the amplitude and frequency rules tend to be offset from each other. This is attributed to the large window of about 50ms used to calculate the fourier transform. This has the effect of causing nasals to be dominated by the frequencies of large neighbouring vowels, and conversely to extend the spectrum of the nasals into adjacent regions of low amplitude. This can be rectified by halving the fourier transform window size (to 25ms) and also applying a Hamming window (cosine bell) or Gaussian window to the fourier transform buffer.

In the meantime an attempt has been made to minimize the effects of this misalignment by extending and moving the segments that fire for the frequency rule before combining them with the amplitude rule.

THE TRAILING NASAL RULE

One further rule was added to the nasal rule described above. This is to recognize the trailing nasal. The frequency part of this rule uses the basic 50 to 400Hz fftband part of the previous rule and extends it slightly to include any small gaps where it may fail to fire. The amplitude part requires abssum to be within a loose range of low values. The result of ANDing these two parts must be at least 100ms long and then be followed by a further period of at least 100ms in which abssum is close to zero. This is achieved using the before connective.

RESULTS

The final rule to recognize nasal consonants then contains one part, using much amplitude information, to recognize trailing nasals, and a second part, based mainly on frequency information, to recognize other nasals.

Results from the final rule for the two sentences spoken by each of the three speakers are given in Table 1. Detection rate (Table 1a) is given as the number of nasal consonants which caused a rule to fire over the total number of nasal consonants in the sample of speech. Misfire rate (Table 1b) is the number of times the rule fired incorrectly over the total number of times that it fired. Adjacent nasals usually fire over a single extended segment. These have been counted as both nasals causing the rule to fire. Occasionally the time over which a rule fires for a single phoneme is broken into several segments. This has occurred, for example, when the amplitude of a long low-amplitude nasal dips just below the cut-off level specified in the rule, causing the region over which the rule fires to break into two segments. This is counted as a single detection. 79% of the nasal consonants were detected, with the nasal rule firing incorrectly 10% of the time. Incorrect firings were due to the end of /u/ in two (three times), the beginning of /w/ in while, the end of /u/ in snow, and the /i/ and the /l/ in dismantling.

	Sentence		
Speaker	n1	n2	Both
1 (female)	18/20	13/18	31/38
2 (female)	14/20	13/18	27/38
3 (male)	17/20	15/18	32/38
3 speakers	49/60	41/54	90/114

Table 1a: Detection Rate

	Sentence		
Speaker	n1	n2	Both
1 (female)	0/18	2/15	2/33
2 (female)	2/16	1/15	3/31
3 (male)	1/8	4/19	5/37
3 speakers	3/52	7/49	10/101

Table 1b : Misfire Rate

DISCUSSION

The results obtained are promising but the derivatives used need to be modified before it is worth refining the rules and testing over large numbers of speakers. The problem with the fourier transform is straightforward. However the difficulties in obtaining a handle on the shape of the waveform are greater.

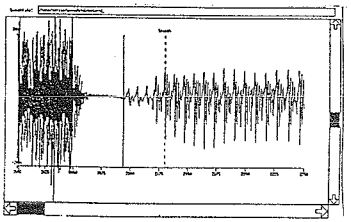

Figure 1: Initial nasal

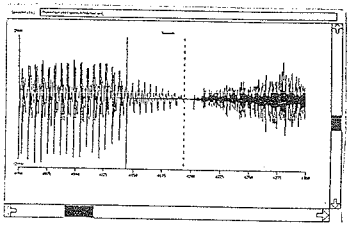

Figure 2: Final nasal

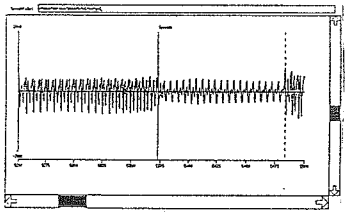

Figure 3: Medial nasal

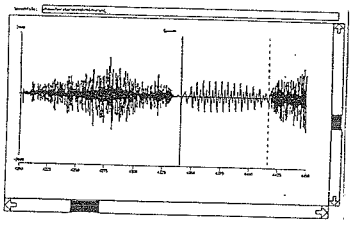

Figure 4: Syllabic nasal

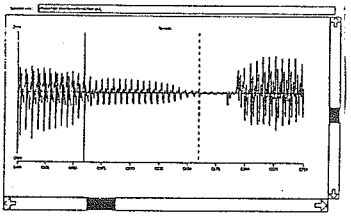

Figure 5: Long-final nasal

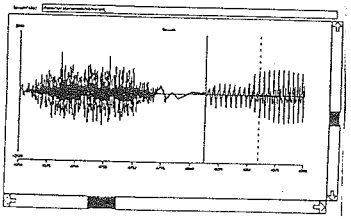

Figure 6: Long-initial nasal

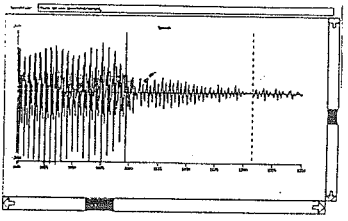

Figure 7: Trailing nasal