Formant-based speech recognition.
A study on a speaker dependent application.

Fredrik Edelstam
freede41@kth.se

October 23, 2014

Abstract

This paper accounts for some implications of using formants for ASR. Also an account is given for a method for how to extract formant features. Finally an experiment is conducted that investigates how well three different sets of formant features perform compared to standard MFCC features in a speaker-dependent application. Feature vectors that also include bandwidth deltas are not considered.

Some implications of using formant features for ASR is that delta features should be included to classify diphthongs and obstruents. Formants can merge and this can be handled with formant tracking. Vocal tract length normalisation should be done to eliminate formant position differences between males and females.

The experiment was done in HTK with 20 training sentences and 10 test sentences, containing sequences of four digits in English. The results of the experiment show that MFCC's outperform formant features with an accuracy of 70 % compared to 45 % for the highest formant feature score, which was achieved by four formants with deltas. The low results for formant features depend in part on the occurrence of the alternative word for zero, "oh", which only occurred once per training and test set.

1 Introduction

1.1 Purpose

This paper accounts for some implications of using formants for ASR. Also an account is given for a method for how to extract formant features. Finally an experiment is conducted that investigates how well three different sets of formant features perform compared to standard MFCC features in a speaker-dependent application. Feature vectors that also include bandwidth deltas are not considered.

1.2 Why use formant features?

For purposes of automatic speech recognition (ASR) it is not desirable to represent the speech spectrum with a complete speech spectrum. Using a 512 point fast Fourier transform (FFT), each feature vector would have a file size of 512 bytes (256 evenly spaced frequency values between zero and the Nyquist frequency, multiplied by two given 16 bits/symbol).

Instead, it is desirable to represent the spectrum in some more compact format. Formants seem like a good candidate for this since two to four formants and their deltas (trajectories) can distinguish between speech sounds. Four formants and four deltas result in a feature vector that is 16 byte long (if 16 bits/symbol). To compute formants, many problems have to be dealt with.

2 Formant implications for ASR

The source for this section is mainly [2]. Also [3] has been used.

Formants are peaks in the speech spectrum due to resonances in the vocal tract (throat/pharynx and mouth/oral cavity) and possibly also the nasal tract. Formants are characterised by their position, amplitude and bandwidth. A naming convention for formants is that they are called F, followed by an index that orders the formants from the low to the high frequencies. All formants do not describe phonemes. Speech sounds above F3 describe the speaker's oral tract. Vowels can be distinguished by F1 and F2 alone. Nasals are distinguished by F1 and F3. Voiced sounds have a lowpass characteristic that weakens formants higher up the spectrum. For this reason, formants above F3 are rarely used.

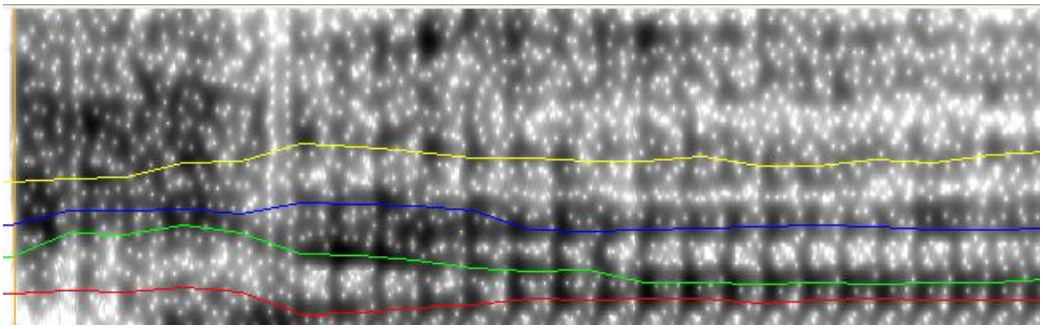

Figure 1: Formant plot of a coarticulation effect from /g/ to /a/, uttered by the author of this paper. F1 and F2 have trajectories that go to lower frequencies.

2.1 Target formants and formant tracks

There is roughly one formant per kHz in the speech spectrum. Some phonemes are characterised by *target formants*. Target formants are assumed in steady-state articulation. Target formants are the formants where the phoneme is most easily identified.

Other phonemes are characterised by the formant tracks. Between target formants, the formants move up and down the spectrum to the next target formant configuration. These up and down movements are called *formant tracks*. *Formant tracking* is any technique used to follow the formant tracks.

Sonorants (vowels and semivowels) and nasals are characterised by target formants. Diphthongs never articulate target formants in steady-state, but always move between the them. Obstruents (plosives and fricatives) are characterised by the coarticulation with neighbouring sonorants. An example illustrates this in figure 1. It shows with a formant plot made in Wavesurfer of the coarticulation of /g/ with /a/. The right context of /g/ is characterised by F1 and F2 downward trajectories.

An implication for formant-based ASR is that delta features should be included.

2.2 Age and gender variation

A problem for formant-based ASR is that the formants of a phoneme varies with age and gender.

The vocal tract can be modelled as a quarter-wavelength resonator. A quarter wavelength resonator is a tube of constant cross-sectional area that is closed at one end and open at the other. The glottis (area between the vocal folds) can be approximated as a closed end since it is relatively small compared to the cross-sectional area of the vocal tract. Quarter wavelength resonators resonate at quarters of wavelengths that are twice the roundtrip distance (four times the length) of the tube.

The length of female and male vocal tracts are different. On average, a male vocal tract is 17 cm long. A female tract is on average 13 cm long. This means that female formants are spaced farther apart than male formants. The vocal tracts of children is even shorter as a function of age. The resonating wavelengths for a vocal tract of constant cross-sectional area is given by the following formula.

$$\lambda_r = \frac{4L}{n} \tag{1}$$

L is the vocal tract length. The relationship below allows a conversion between wavelength and frequency.

$$f = \frac{c}{\lambda} \tag{2}$$

c is the speed of sound. What the formula above does, dividing the number of metres that sound travels in air in a second by the number of metres that one wavelength occupies, is the same as counting the number of wavelength per second. Assuming that c is 340 m/s, this means that formants are placed at odd multiples of 500 Hz in a constant cross-section vocal tract.

$$\frac{340m}{4 \times 0.17m \times s} \approx 500Hz \tag{3}$$

The spacing of female formants is calculated below.

$$\frac{340}{0.52} \approx 650Hz \tag{4}$$

The implication for formant-based ASR is that a speaker-independent application should try to normalise the vocal tract length, if it is possible to estimate the vocal tract length just from speech data.

2.3 Merging formants

Only the schwa (pronounced as the i in *bird*) is pronounced with a vocal tract of constant cross-sectional area. All the other phonemes are deviations from this. The phonemes deviate because of varying sizes of the throat and the mouth. When the throat transitions into the mouth abruptly, this causes problems for formant-based ASR.

This is also when the vocal tract shape deviates most from the schwa. At abrupt transitions, the throat and mouth produce formants relatively independently. This means that formants almost merge. There will still be some dependency, so formants never come closer than about 200 Hz. It will be difficult for formant-based speech recognisers to determine whether a frequency band contains one or two formant frequencies. The solution is to track the formants.

The phonemes with the most extreme deviations are called the cardinal vowels. The cardinal vowels can be different for different languages, but mostly they are /a/ as in *hot*, /u/ as in tool and /i/ as in feet. F1 and F2 almost merge for /a/ and /u/. F2 and F3 almost merge for /i/. The liquids /r/ and /l/ also have close F2 and F3.

F1 describes the resonance of the throat. When the tongue is high, the throat is longer and consequently F1 is lower. F1 ranges from 300 Hz (high/close vowels) to 800 Hz (low/open vowels).

F2 describes the resonance of the mouth. When the tongue is back, F2 is low. F2 ranges from 700 Hz (back vowels) to 2200 Hz (front vowels). This means that a high back vowel has a low F1 and also a low F2.

F3 decribes how curled the tongue is. F3 usually ranges from 1800 (retroflex) to 2800 (high-front vowels).

3 How to compute formant features

Bohm and Nemeth [1] compute formants by solving for roots in the LPC spectra and selecting formants from the roots by using constraints.

Formants are computed by finding local maxima in some representation of the spectrum. It is better to use the linear prediction coefficient (LPC) spectrum than the FFT spectrum. The LPC spectrum is specialised at estimating the spectral peaks. Its resolution can also be increased. The LPC spectrum is given by the reciprocal of the LPC polynomial:

$$H(z) = \frac{1}{1 - \sum_{k-1}^{p} \alpha_k z^{-k}} \tag{5}$$

p is the order of prediction and it tells how many poles that the spectrum includes. As a rule of thumb, 2 poles per formant are used. Computing six formants means that a prediction order of 12 should be used.

3.1 Constraints

To find the formants, the roots for the LPC polynomial are solved for. Some of the roots are possible formants, *formant candidates*. To decide which roots could be formants, the bandwidths of the roots are computed. The bandwidth is the width of the frequency band where the pole is less than 3 dB weaker than its centre frequency. It is computed with the following formula:

$$B_i = \frac{f_s}{\pi} ln(^1/_{r_i}) \tag{6}$$

Bandwidths lower than 50 Hz and wider than 300 Hz indicate that the root is caused by background noise or a strong harmonic.

Also, roots below the fundamental frequency are discarded (provided that the fundamental frequency has been calculated, for example with the autocorrelation method).

To find the formants from the formant candidates, formant tracking is done. Each formant candidate is mapped to the nearest formant in the previous frame. Tracks that run in parallel close to each other are merged into one track. This means that two formant candidates are treated as one. Also, tracks that do not collide at all or only collide minimally are treated as one. The authors do not say what happens in the case that there are more formant candidates than tracks in the past frames. Perhaps this problem does

not occur within state boundaries. The authors say that abrupt changes in formant positions are a problem if transcriptions are not available.

4 Method

In this study, three different types formant feature vectors are tested on speech data from one speaker in HTK. The speech data consists of sequences of four digits in English. The feature vectors consists of:

1. four formants
2. four formants and four deltas
3. four formants, four deltas and four bandwidths

The speech data is also trained and tested with the standard ASR feature vector (39 MFCC including deltas and accelerations) and compared to the formant feature results.

4.1 Training and testing material

All three feature vectors were trained on 20 pre-recorded utterances and tested on 10 pre-recorded utterances. All instances were spoken by the same person. The training material and the testing material were recorded on different days. The utterances were recorded with a sampling frequency of 44100 Hz and a representation of 16 bits/symbol. The microphone used for recording belongs to the headset Logitech PC Headset 120. The recording was done on a laptop running Audacity.

The wave files were orthographically transcribed manually in Wavesurfer and saved with the .lab extension, which is a format accepted by HTK.

A formal grammar was constructed that consisted of the numbers zero to nine in English including an alternative way to say zero (*oh*). This utterance occurred once in the training data.

4.2 Feature extraction

Wavesurfer was used to extract four formants and their bandwidths. The data was saved as text files. The text files were read into Matlab where the

formant deltas were computed. The feature vector representation of each utterance was saved as a separate text file.

Since these files do not describe waveforms they had to be converted into binary files so that HTK could read them. The text files were converted using a C++ program Windows written by Kalin Stefanov at KTH. The files were saved in the .ext format.

4.3 Procedure

The shell scripts from the lab in the course DT2118 were used to train and test the formant features. The selected feature was USER. The sourceformat specified in the configuration file for the USER parameter kind was set to HTK. This is the binary format. Targetkind and sourcekind in the configuration file were set to USER. The variables relating to signal processing such as windowsize were deleted since the processing had already been done in Wavesurfer.

In the tables below, H refers to the number of correctly recognised words, D refers to deletions, S refers to substitutions, I refers to insertions and N refers to the total number of words in the training data. %Corr refers to the percentage of correctly recognised words and is calculated as $^{H}/_{N}$. Acc is the accuracy and it is calculated as $\frac{(H-I)}{N}$

5 Results

5.1 Formants only

```
====================== HTK Results Analysis =======================
  Date: Sun Oct 19 16:55:25 2014
  Ref : workdir/all_word.mlf
  Rec : results_USER/recout_test.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=37.50, Acc=32.50 [H=15, D=8, S=17, I=2, N=40]
------------------------ Confusion Matrix -------------------------
       E   F   N   O   O   S   T   Z
       I   O   I   H   N   I   H   E
       G   U   N       E   X   R   R
       H   R   E               E   O
       T                       E      Del [ %c / %e]
EIGH   4   0   0   0   0   1   0   0    0 [80.0/2.5]
FIVE   2   1   0   0   0   0   0   0    1 [ 0.0/7.5]
FOUR   1   1   0   2   0   0   0   0    0 [25.0/7.5]
NINE   1   1   1   0   0   0   0   0    1 [33.3/5.0]
  OH   0   0   0   1   0   0   0   0    0
 ONE   0   0   1   0   1   0   0   0    2 [50.0/2.5]
SEVE   1   0   1   0   0   0   0   0    2 [ 0.0/5.0]
 SIX   1   0   0   0   0   2   0   0    0 [66.7/2.5]
THRE   1   0   0   0   0   0   3   0    0 [75.0/2.5]
 TWO   0   0   0   3   0   0   0   0    1 [ 0.0/7.5]
ZERO   0   0   0   0   0   0   0   2    1
Ins    2   0   0   0   0   0   0   0
===================================================================
```

The results for the first feature vector shows that "five", "seven" and "two" have never been detected. There were two insertions, eight deletions and 17 substitutions. The greatest confusion was between "two" and "oh", which were confused three times. Second greatest confusion was between "four" and "oh", and "five" and "eight", which were confused two times. "Five" was confused with "eight" two times. The percentage of correct words was 37.5 %.

5.2 Formants and deltas

```
====================== HTK Results Analysis =======================
  Date: Sun Oct 19 16:59:01 2014
  Ref : workdir/all_word.mlf
  Rec : results_USER/recout_test.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=45.00, Acc=35.00 [H=18, D=6, S=16, I=4, N=40]
------------------------ Confusion Matrix -------------------------
       E   F   O   O   S   S   T   Z
       I   O   H   N   E   I   H   E
       G   U       E   V   X   R   R
       H   R           E       E   O
       T               N       E       Del [ %c / %e]
EIGH   4   0   0   0   0   0   0   0     1
FIVE   1   0   1   0   0   0   0   0     2 [ 0.0/5.0]
FOUR   1   1   2   0   0   0   0   0     0 [25.0/7.5]
NINE   3   0   0   0   0   0   0   0     1 [ 0.0/7.5]
  OH   0   0   1   0   0   0   0   0     0
 ONE   0   0   0   4   0   0   0   0     0
SEVE   1   0   0   0   1   0   0   1     1 [33.3/5.0]
 SIX   1   0   0   0   0   1   0   0     1 [50.0/2.5]
THRE   1   0   0   0   0   0   3   0     0 [75.0/2.5]
 TWO   0   0   4   0   0   0   0   0     0 [ 0.0/10.0]
ZERO   0   0   0   0   0   0   0   3     0
Ins    3   0   1   0   0   0   0   0
===================================================================
```

The results for the second feature vector shows that "five", "nine" and "two" have never been detected. This time "seven" has been detected. There were four insertions, six deletions and sixteen substitutions. The greatest confusion was again between "two" and "oh", which were confused four times. The second greatest confusion was between "nine" and "eight", which were confused three times. "Four" was also confused with "oh" two times. The percentage of correct words was 45.0 %.

5.3 Formants, deltas and bandwidths

```
====================== HTK Results Analysis =======================
  Date: Sun Oct 19 17:04:23 2014
  Ref : workdir/all_word.mlf
  Rec : results_USER/recout_test.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=10, N=10]
WORD: %Corr=42.50, Acc=15.00 [H=17, D=3, S=20, I=11, N=40]
------------------------ Confusion Matrix -------------------------
       E   F   N   O   O   S   S   T   Z
       I   O   I   H   N   E   I   H   E
       G   U   N       E   V   X   R   R
       H   R   E           E       E   O
       T                   N       E      Del [ %c / %e]
EIGH   5   0   0   0   0   0   0   0   0    0
FIVE   2   0   0   0   0   0   0   0   0    2 [ 0.0/5.0]
FOUR   1   0   0   3   0   0   0   0   0    0 [ 0.0/10.0]
NINE   0   0   1   2   0   0   0   0   0    1 [33.3/5.0]
  OH   0   0   0   1   0   0   0   0   0    0
 ONE   0   0   0   0   4   0   0   0   0    0
SEVE   1   1   0   0   1   1   0   0   0    0 [25.0/7.5]
 SIX   1   0   0   1   0   0   1   0   0    0 [33.3/5.0]
THRE   1   0   0   0   0   0   0   3   0    0 [75.0/2.5]
 TWO   0   0   0   4   0   0   0   0   0    0 [ 0.0/10.0]
ZERO   0   0   0   2   0   0   0   0   1    0 [33.3/5.0]
Ins    3   1   2   5   0   0   0   0   0
===================================================================
```

The results for the first feature vector shows that "five" and "two" have never been detected. This time, "nine" and "seven" were detected, however. There were eleven insertions, three deletions and twenty substitutions. The greatest confusion was again between "two" and "oh", which were confused four times. The second greatest confusion was "four" and "oh", which were confused three times. The percentage of correct words was 42.5 %.

5.4 MFCC

```
====================== HTK Results Analysis =======================
  Date: Sun Oct 19 16:49:16 2014
  Ref : workdir/all_word.mlf
  Rec : results_MFCC_0_D_A/recout_test.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=20.00 [H=2, S=8, N=10]
WORD: %Corr=70.00, Acc=47.50 [H=28, D=0, S=12, I=9, N=40]
------------------------ Confusion Matrix -------------------------
       E   F   F   N   O   O   S   S   T   T   Z
       I   I   O   I   H   N   E   I   H   W   E
       G   V   U   N       E   V   X   R   O   R
       H   E   R   E           E       E       O
       T                       N       E           Del [ %c / %e]
EIGH   4   0   0   0   0   0   0   0   0   1   0    0 [80.0/2.5]
FIVE   0   4   0   0   0   0   0   0   0   0   0    0
FOUR   0   0   4   0   0   0   0   0   0   0   0    0
NINE   0   3   0   1   0   0   0   0   0   0   0    0 [25.0/7.5]
  OH   0   0   0   0   0   0   0   0   0   1   0    0 [ 0.0/2.5]
 ONE   0   1   0   0   0   3   0   0   0   0   0    0 [75.0/2.5]
SEVE   0   0   0   0   0   0   3   0   0   0   1    0 [75.0/2.5]
 SIX   1   0   0   0   0   0   0   2   0   0   0    0 [66.7/2.5]
THRE   0   0   1   0   0   0   0   0   2   0   1    0 [50.0/5.0]
 TWO   0   1   0   0   0   0   0   0   1   2   0    0 [50.0/5.0]
ZERO   0   0   0   0   0   0   0   0   0   0   3    0
Ins    2   2   0   1   2   0   0   0   0   1   1
===================================================================
```

The results for the MFCC vector shows that all words were detected. There were nine insertions, no deletions and twelve substitutions. The greatest confusion was between "nine" and "five", which were confused three times. Other confusions only happened once. The percentage of correct words was 70.0 %.

6 Discussion

The results show that there were many substitutions. There were more substitutions using formant features than MFCC. The most substitutions happened when bandwidths were also used.

The digits that the recogniser failed to recognise were "seven", "five" "two" and "nine". "Five" and "two" were never detected for any formant feature set.

In the first feature set, only formants were used. Only sonorants and nasals are characterised by formants, so this implies that words containing the same set of sonorants were substituted. "Three", "zero", "seven" and "six" all contain sonorants. This feature set was good at detecting "three" and "zero", but it failed to detect "seven". No other digit contains /EH/, so it should have been easy to discriminate "seven".

Notably, "two" and "four" were often substituted with "oh". The confusion matrix says there were no deletions and no insertions at these points. The words sound similar, but they contain different diphthongs: /UW/ in case of "two", /AO/ in the case of "four" and /OW/ in case of "oh". There was only one training instance of "oh" It seems reasonable to believe that this makes it more difficult for the recogniser to discriminate "oh". It did detect the instance of "oh" in the test set, however.

In the second feature set, performance rose with 7.5 percentage units. This makes it reasonable to conclude that some of the errors in the previous set depended on poor discrimination of diphthongs and obstruents. Again, "two" and "four" were substituted with "oh". It seems reasonable to draw the conclusion that the models are similar. This might depend on the speaker's pronunciation at the training occasion.

For this set, "nine" and "eight" were confused. The confusion matrix says that "eight" was inserted three times. If "eight" was inserted after "nine" then this could depend on the formant track between /EY/ and /T/ being similar to the formant track between /AY/ and /N/. It is not safe to draw any conclusion other than that the inclusion of "oh" affected the results again.

Performance dropped 2.5 percentage units with the inclusion of bandwidths. The feature vector that includes bandwidth has the highest number of substitutions. This implies that bandwidths are not so good at discriminating speech sounds. Again, "two" and "four" were confused with "oh".

7 Conclusion

This paper has outlined some implications for ASR when using formant features. Formant features vary over gender and age so this has to be addressed during feature extraction. Target formants only characterise sonorants and nasals. Obstruents are characterised by formant tracks. Therefore delta features should be included in a formant feature vector. Two formants can merge and look like one formant. This problem can be addressed in formant-based recognisers by formant tracking.

A feature extraction method for formants is presented that is used by Bohm and Nemeth [1]. This method extracts formants from the LPC spectra by calculationg the roots of the LPC polynomial. The method removes roots that cannot be formants by setting constraints on the bandwidths and the formant position. Formants are chosen from the formant candidates by formant tracking.

An experiment in HTK was performed on sequences of four digits. The results showed that there were many substitutions. Performance was best for formant features that included deltas. Performance dropped when including bandwidths. Including "oh" once in training and test set seems to have lowered the results. Even under this condition MFCC performed a lot better than formants with delta features, recognising the correct word 70 % of the time compared to 45 % of the time for the best formant feature vector (formants and deltas).

References

[1] Bhm, T. and Nmeth, G. (2007). "Algorithm for formant tracking, modification and synthesis." In *HRADSTECHNIKA* (Vol. 62, pp. 15-20)

[2] O'Shaughnessy, D. (2008). "Formant estimation and tracking" In *Handbook of speech processing*, Eds. Benesty, J., Sondhi, M. M., and Huang, Y. pp. 213-228

[3] Hung, X., Acero, A. and Hon, H. (2001). *Spoken language processing. A guide to theory, algorithm and system development* USA: Prentice Hall