Qiong Hu, Korin Richmond, Junichi Yamagishi, and Javier Latorre. An experimental comparison of multiple vocoder types. In 8th ISCA Workshop on Speech Synthesis, pages 155-160, Barcelona, Spain, August 2013.

Table 1: *Summary of selected vocoders (k: number of sinusoids per frame, HTS: the suitability for HTS modelling).*

Name	Vocoder	HTS	Parameters per frame
MGC	Mel - generalised cepstral vocoder	Yes	MGC: 24 + F0: 1 , Pulse plus noise excitation
SF	STRAIGHT with full band mixed excitation	No	Aperiodicity:1024, spectrum: 1024+ F0:1, Multi- band mixed excitation
SC	STRAIGHT-MGC with critical band mixed excitation	Yes	Band aperiodicity: 25 + MGC :39 + F0: 1, Multi-band mixed excitation
Glot	Glottal vocoder	Yes	F0:1, Energy:1, HNR: 5, Source LSF: 10, Vocal tract LSF: 30, natural pulse
DSMR	MGC vocoder with DSM-based residual	Yes	MGC: 30 + F0:1, DSM for residual excitation
HM	Harmonic model	No	2*k harmonics + F0:1, Harmonic excitation
HMF	Harmonic with fixed dimension	No	2*k harmonics + F0: 1, Harmonic excitation
HNM	HNM-MGC vocoder	Yes	MGC:40 + F0:1, Multi- band excitation, Maximum voiced frequency
aHM	Adaptive harmonic model	No	2*k + F0:1, Harmonic excitation
OS	Original speech		

mel-cepstral coefficients and F0 [9], adaptive harmonic vocoder [6], harmonic vocoder [8], harmonic vocoder with fixed parameters were selected. For the source-filter vocoders,the deterministic plus stochastic model for residual(DSMR) vocoder [4], the mel-generalized cepstral vocoder, glottal vocoder [5], and STRAIGHT [3] with both full-band and critical-band based mixed excitation [10] were chosen for comparison.

2.1. Mel-generalized cepstral vocoder (MGC)

Here, a simple pulse/noise excitation is used for the MGC vocoder. Although straightforward, this excitation model cannot fully represent natural excitation signals and often generates "buzzy" speech. Different types of coefficients may be used to represent the spectrum. Mel-cepstra are often used, providing a good approximation to the human auditory scale of speech. Here, we use the Mel-Generalised Log Spectral Approximation (MGLSA) digital filter to filter the excitation signal to synthesise speech. We use the same parameter value in [4] α=0.42 and γ=-1/3 for MGC extraction.

2.2. STRAIGHT with full-band mixed excitation (SF)

STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of Weight Spectrum) [3] was developed to better remove the periodicity effects of F0 on extracting the vocal tract spectral shape. For spectral envelope extraction, both F0 adaptive spectral smoothing and compensatory time windows are used to transfer the time frequency-smoothing problem to frequency domain. Aperiodicity of the signal is computed as the difference between the upper and lower envelope of the spectrum. For voiced frame, noise is calculated by modulating the randomness of the phase component according to aperiodicity. Finally, all parameters are sent to a minimum-phase filter with group delay phase manipulation to synthesise speech.

2.3. STRAIGHT mel-generalised cepstal vocoder with critical band mixed excitation (SC)

Although STRAIGHT uses both aperiodicity and F0 adaptive spectral smoothing to solve the "buzzy" problem, the number of parameters for both the spectrum and aperiodicity components is the same size as the FFT length used, which is not suitable for statistical modelling. [10] proposed to use other lower dimensional parameters, such as Mel-generalized Cepstral Coefficients or Line Spectral Pairs to represent the spectrum instead. Here, in order to compare with other vocoders with similar spectral parameters, the Mel-generalised cepstral is chosen as the intermediate spectral parameterization. Aperiodicity parameters are also compressed by averaging the whole points to 25 sub-bands. The same type of filter is chosen as used in the STRAIGHT vocoder above.

2.4. Glottal vocoder (Glot)

[5] proposed a method to represent the glottal pulse signals instead of using a pulse-train excitation to represent the voiced excitation. For voiced speech frames, Interactive Adaptive Inverse Filtering (IAIF) is used to separate the glottal source from the vocal tract so that both the vocal tract and source signal may be accurately estimated. For unvoiced frames, conventional inverse filtering is applied. Other parameters, such as energy and harmonic-to-noise ratio (HNR), are calculated so as to weight the noise component of the source. During synthesis, a prestored library pulse is selected and interpolated to match the target F0. The glottal spectrum, HNR and energy also have to be set to match the target. Finally, a vocal tract filter as derived from analysis part is applied to the excitation to generate the speech signal.

2.5. MGC vocoder with DSM-based residual (DSMR)

In [4], a MGC vocoder with Deterministic plus Stochastic Model for residual signal is proposed. The residual signal is first obtained by applying inverse filtering using mel-generalised cepstrum filters. Then, a Blackman window, centred on glottal closure instants and of length equalling two F0 periods, is applied to obtain pitch-synchronous residual frames. In order to model these, they are first length normalised, then the deterministic component at the lower frequencies is decomposed using Principal Component Analysis (PCA) to obtain the first eigen residual. The energy envelope and an autoregressive model are used for the stochastic component. During synthesis, both these parts are resampled to match the target pitch to produce the new residual signal, which is used to drive a MGLSA filter to generate speech, so it is not strictly a sinusoidal vocoder.

2.6. Harmonic vocoder (HM)

Although real amplitude for the sinusoids were used for calculating parameters in [11], complex amplitudes proposed by [8], estimated by an algorithm operating in the time domain, are used in our experiment here, as it is easier to deal with the phase information (e.g. we can avoid problems such as phase unwrapping). For voiced frames, we calculate the complex amplitude by minimising the error between the original and estimated speech signals. The number of harmonics k in each frame is dictated by Fs/F0 (Fs: sampling frequency, F0: fundamental frequency). For unvoiced parts, Karhunen-Loeve expansion [12] shows we can use the same analysis as for voiced frames. We suppose that the frequency are close enough and set the F0 as 100Hz under the window length of 20ms to make the power spectrum change more slowly. From the complex amplitudes for a sequence of frames, we use the standard overlap and

add technique to re-synthesis speech.

2.7. Harmonic vocoder with fixed dimension(HMF)

From the description of the Harmonic Vocoder in the previous section, note the number of complex amplitude values in each frame varies depending on F0. This varying number of parameters is not suitable to combine with HTS. So, we also include a variant of the previous Harmonic vocoder in our experimental comparison that uses a fixed number of parameters per frame, which is labelled the "HMF" vocoder. To fix the number of harmonics, one option is to use those harmonics in at lower frequencies and add noise at higher frequencies. However, dividing the spectrum into two in this way would be rather arbitrary. For unvoiced speech in the "HM" vocoder, the number of harmonics in each frame is fixed, even though there may be no harmonics in fact. Similarly, here we suppose that the number of harmonics is the same as used for unvoiced parts irrespective of whether there are harmonics at higher frequencies or not.

2.8. HNM-MGC Vocoder (HNM)

A harmonic/stochastic waveform generator is presented by [9]. This method is based on the decomposition of the speech frames into a harmonic part and stochastic part and uses MGC, F0 and maximum voiced frequency (MVF) as an intermediate parameterization. This vocoder is thus suitable for statistical modelling with a fixed frame size. For voiced frames,the entire spectral envelope may be obtained by interpolating the amplitudes at harmonic points. Cepstral coefficients are obtained from the log spectrum and then they are reduced in number [2] and warped to the mel scale. Unvoiced part is just analysed through a fast Fourier transformation (FFT) and no stochastic part is assumed during analysis. MVF is caculated based on sinusoidal likeness measure. During synthesis, the cepstral envelope is resampled according to the harmonic points. Noise component is obtained by sampling the cepstral envelope at frequency above MVF. Minimum phase is using here.

2.9. aHM-AIR vocoder (aHM)

For the "HMF" and "HM" vocoders, we represent the whole band with harmonics alone. In principle, though, small errors in F0 value could cause large mismatch error in the higher frequencies. In order to solve this problem, [6] proposes a full-band adaptive harmonic vocoder without using any shaped noise. For analysis, it uses an Adaptive Iterative Refinement (AIR) method and an adaptive Quasi-Harmonic vocoder (aQHM) as an intermediate model to iteratively minimise the mismatch of harmonic frequencies while increasing the number of harmonics. Then, instantaneous amplitude and phase values may be obtained by interpolation. During synthesis, the aHM-AIR vocoder could represent the same structure by using only F0 rather than a frequency value at each analysis instant.

3. Experiment

3.1. Subjective analysis

Our approach to comparing and analysing the vocoders summarised in Section 2 relies upon multi-dimensional scaling (MDS)[14]. This technique aims to map points within a high dimensional space to a lower dimensional space while preserving the relative distances between the points. We can exploit this to visualise relative distances between the vocoders which indicate similarity in terms of perceptual quality. Listeners are asked to judge whether a given pair of stimuli are the same in terms of quality or different. Comparing a number of stimuli synthesised by all vocoders in this way, we obtain a matrix of inter-vocoder distance scores. This high-dimensional similarity matrix can be reduced to a 2- or 3- dimensional space to visualise vocoder similarities in terms of listener perception. The "Classical MDS" variant is used here, as we are comparing the Euclidean distance between each vocoder. Note we have found the natural speech is perceived as quite different from the vocoded speech, so including natural stimuli can heavily distort the relative distances between each vocoder if included. Therefore, we have omitted it from our MDS analysis. Instead, preference tests are subsequently used in order to compare the quality of each vocoder against the original speech.

Table 2: *Parameters for each section*

Section	Speaking style	Questions	ratio
1	Normal	Similarity	0.7943
2	Lombard	Similarity	0.7760
3	Normal	Preference	0.7500
4	Lombard	Preference	0.7451

In the test, every vocoder is compared pairwise with all others, giving a 9*9 similarity matrix. Phonetically balanced speech data from a UK male speaker is used for copy synthesis with each vocoder. The sampling rate is 16kHz. A total of 32 normal speaking style sentences and another 32 different sentences with Lombard speaking style are used. Several samples are available on the webpage (http://homepages.inf.ed.ac.uk/s1164800/vocoder_com.html). For each comparison unit and each listener, sentences are randomly selected for the matrix. So, all possible sentences could be heard for each comparison to mitigate sentence-dependent effects. Forty one native English speakers participated in the listening test, conducted in perceptual sound booth with headphones. Moreover, we suspect that the questions used for the listening test (same/different or better/worse/same) and the type of sentences (Lombard or Normal) could affect the MDS result as well. So, four sections are designed to test for this effect. A summary of the speaking styles, questions for comparing sentences and the eigenvalues ("ratio") for the first two dimensions found by MDS analysis are listed in Table 2.

The two-dimensional MDS spaces for the four test sections are shown in Figure 3. At first sight, it seems the locations of the vocoders differ in each section. However, by comparing the four MDS figures, we can see that although the absolute x- and y-coordinates for each point may vary, the relative positions of each vocoder are similar. The approximate consistency between the 4 different test sections indicates the relative layout of the vocoders observed is to some extent general, and that sufficient and adequate test stimuli have been selected, for example.

Next, we aim to analyse and interpret the relative layout of the vocoder points in the MDS space. Different speaking and question styles are used in each test section, and so we use Analysis of Variance (ANOVA) to ascertain whether these factors explain the variations observed. The results of both one-way and two-way ANOVAs are shown in Table 3. For the one-way method, the F-values for both speaking and question style for MDS are high. Meanwhile, both significances are less than 5 percent, which means these two factors greatly affect listener judgement. The two-way ANOVA indicates there is no significant interaction between the effects of speaking style and question type on listener judgement. We conclude therefore that speaking style and question format to some extent explain why each section map differs. Furthermore, in Table 2, note the ratio for the "same/different" question type is higher than that ob-

Table 3: *ANOVA for speaking style and question type*

Type	Anova	F value	Significance
One-way	Data~Style	6.7775	0.00993
One-way	Data~ Question	18.659	2.471e-05
Two-way	Data~Style*Question		
	Style	7.3651	0.007243
	Question	19.1647	1.949e-05
	Style:Question	0.0006	0.980126

tained used the 3-way "better/worse/same" question type. We believe therefore the first question type may yield more dependable results. So, for objective analysis, only section 1 and 2 are used for Normal speech and Lombard speech separately.

Although proximity in the MDS map can be interpreted as similarity, the relationship between the vocoders is not yet necessarily clear, so it would be more obvious to merge similar vocoders together. Thus, based on the 9*9 matrix of Euclidean distance between each vocoders, we use K-means clustering to identify emergent groupings. The "Silhouette" value [13] for varying numbers of clusters is computed, and the highest value is taken to indicate the optimum cluster number. The result for each test section is shown in Figure 4. The MDS results show that the SC, SF, MGC and Glot vocoders are very close to each other, indicating listeners find they sound similar to one another. A similar situation is observed for the DSMR and HNM vocoders, and for the aHM and HM vocoders. The clustering result in Figure 4 is consistent with this. In test section 1, except DSMR which uses DSM for residual siganl but is still based on source-filter model, vocoders in cluster two (in red) all use harmonics to describe speech. It is interesting that they all cluster separately from cluster one (in blue), where the vocoders belong to the traditional source-filter paradigm. More specifically, SC is merely a reduced dimension version of SF. Meanwhile, the intermediate parameters transferred from spectrum is the Mel Generalized Cepstrum, so it is also reasonable for MGC vocoder to be close to SF and SC. For other test sections, the situation is similar except for the relative change of the HM and HMF vocoders. Thus, we conclude that in terms of quality, the sinusoidal vocoders in this experiment sounds quite different from source filter vocoders, and there may be other reasons for DSMR clustering together with sinusoidal vocoders.

Having established similarities between vocoders, we also assess their relative quality compared to natural speech. A preference test is conducted for this purpose. Thirty two normal sentences and another 32 Lombard speech are surveyed separately. The same 41 native listeners participated in this test to give their preference in term of quality. The results given in Figure 1 show that the sinusoidal vocoders give relatively good quality. To further analyse the robustness of each vocoder for modelling both Normal and Lombard speech, the difference in preference scores between these 2 speech styles is presented in table 4. As we can see, in general, sinusoidal vocoders like HMF, HM and aHM give much less variable performance than the source/filter vocoder type. Interestingly, the SF vocoder gives stronger performance in terms of listener preference for Lombard speech than it does for normal speech in Figure 1. The reason for this is the subject of ongoing research.

3.2. Objective analysis

In this section, we explore why the vocoders cluster together as observed and what potential factors underpin listener judgements. A range of standard acoustic objective measures are calculated:

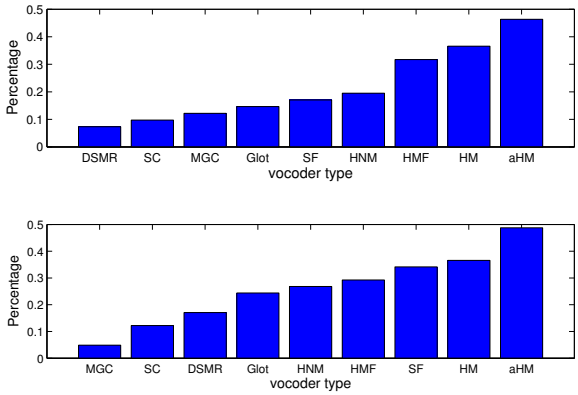


Figure 1: *Preference Test Result (up: Normal, down: Lombard)*

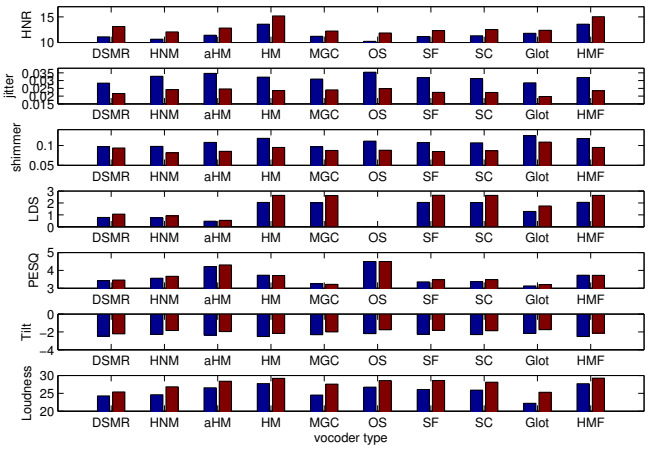


Figure 2: *Objective value result (blue: Normal , red: Lombard)*

- HNR (Noise Harmonic Ratio)
- Jitter
- Shimmer
- LDS (Log distance of spectra using FFT)
- PESQ (Perceptual Evaluation of Speech Quality)
- Spectral Tilt
- Loudness (Based on Model of ISO 532B)

The mean values for these acoustic measures are shown in Figure 2. Unfortunately, we can find no obvious relationship between these measures and the distances between the different vocoders. We attempt to interpret the significance of the MDS map axes by using linear regression and stepwise regression between the two axes and the given acoustic measures. As space is limited here, only the measure most highly correlated with the axes is listed in Tables 5.

As Table 5 shows, the significance of the correlation between PESQ scores with one axis of the MDS map is strong. In fact, combined with Figure 2, we can track vocoder quality through the axis value in MDS to a certain degree. For example, in test section 1 for normal speech, lower x-coordinates indicate higher quality in the vocoder. A similar situation applies

Table 4: *Vocoder preference stability result (Lombard preference value minus that for normal speech)*

vocoder type	DSMR	HNM	aHM	HM	MGC	SF	SC	Glot	HMF
preference vaule (Lombard - Normal)	0.0976	0.0732	0.0244	0	-0.0732	0.1707	0.0244	0.0976	-0.0244

Table 5: *linear regression result.*

linear regression	Significance	R squared
Section1_x~PESQ	0.00174	0.7746
Section2_x~PESQ	0.00991	0.6372

to Lombard speech in test section 2. The aHM vocoder has the best quality, followed by the HM vocoder. Note, though, that neither of these are currently suitable for statistical modelling. For the source-filter vocoders, the Glot, SF and SC ones all sound much better than MGC, and they are suited to modelling as well. Of the sinusoidal vocoders, not only are the HNM and DSMR vocoders suitable for modelling, but also appear to give good vocoded speech quality. The HMF vocoder also appears effective for producing speech with a fixed number of parameters. Finally, we consider which acoustic feature may be most related with other MDS axis. Unfortunately, there is no apparent pattern between any acoustic measure and the axis in the stepwise multi-linear regression. Therefore, we conclude that the listener perception judgements may be a more complex combination of multiple potential features.

4. Discussion and conclusion

This paper examines a broad range of vocoders and presents an experimental comparison to evaluate their relationship and potential factors that correlate with perceived vocoder quality. Both Lombard and normal read speech are used as stimuli produced by copy synthesis with each vocoder. MDS is conducted on the listener responses to analyse similarities in terms of quality between the vocoders. Four combinations of speaking style and listening test question format are tested. ANOVA results shows both speaking style and question format greatly affect listener judgements. For the preference question type, the eigenvalues for the first two dimensions in MDS space are somewhat reduced. Thus, we deem the similarity question type is more suitable for MDS analysis, and Lombard and Normal speech are surveyed separately in the subsequent analysis. Comparing preference test results for Normal and Lombard speech, we also find that sinusoidal vocoders give more consistent performance than source filter vocoders.

To analyse their potential relationship in more depth, K-means clustering is applied to the listener similarity judgment matrix and combined with the MDS results. We find in terms of quality, the sinusoidal vocoders cluster separately from the source filter vocoders. Thus, we conclude that sinusoidal vocoders are perceptually distinguishable from source filter ones. The preference test comparisons with the natural stimuli presented here indicate sinusoidal vocoders can give superior vocoded speech quality. In order to interpret the axes of the obtained MDS space, a several objective acoustic measures are tested for correlation with the MDS space axes. Linear regression result shows that one axis is related with quality. However, no obvious acoustic measure could be found to explain the other axis of the two dimensional MDS space, which we interpret as implying that human perception of vocoded speech quality may combine multiple factors.

5. Acknowledgements

This research is supported by Toshiba. The authors also greatly appreciate help from Gilles Degottex (University of Crete), Tuomo Raitio (Aalto University), Thomas Drugman (University of Mons) and Daniel Erro (University of the Basque Country) by generating samples from their vocoder implementations.

6. References

[1] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0." Proc. of Sixth ISCA Workshop on Speech Synthesis. 2007, pp. 294–229.

[2] K. Tokuda, T. Kobayashi, T. Masuko and S. Imai, "Melgeneralized cepstral analysis — A unified approach to speech spectral estimation," Proc. ICSLP'94, pp.1043– 1046, Yokohama, Japan, Sep. 1994

[3] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, vol.27, pp.187-207 (1999).

[4] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," IEEE Trans. Audio, Speech and Language Processing, vol. 20, no. 3, pp. 968–981, March 2012.

[5] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," IEEE Trans. on Audio, Speech, and Lang. Proc., vol. 19, no. 1, pp. 153–165, Jan. 2011.

[6] G. Degottex and Y. Stylianou. "A Full-Band Adaptive Harmonic Representation of Speech."In Proc. Interspeech, Portland, USA. ISCA, September 2012.

[7] M. Airaksinen. "Analysis/Synthesis Comparison of Vocoders Utilized in Statistical Parametric Speech Synthesis." Master thesis, Aalto University, November 2012

[8] Y. Stylianou, "Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification", Ph.D. thesis, Ecole Nationale Superieure des Telecommunications, January 1996.

[9] D. Erro, I. Sainz, E. Navas, I. Hernaez, "Improved HNM-based Vocoder for Statistical Synthesizers", Proc. Interspeech, pp. 1809-1812, Florence, August 2011.

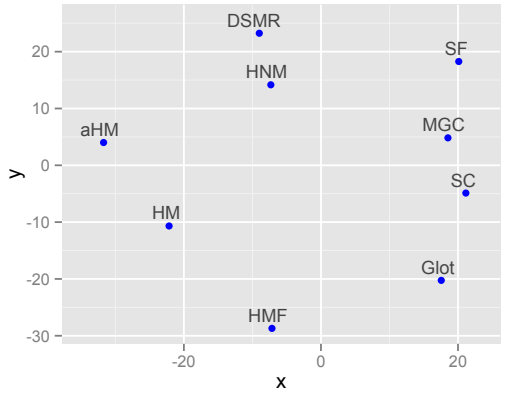
[10] H. Zen, T. Toda, M. Nakamura, and T. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," IEICE Trans. Inf. & Syst., vol. E90-D, no. 1, pp. 325–333, 2007.

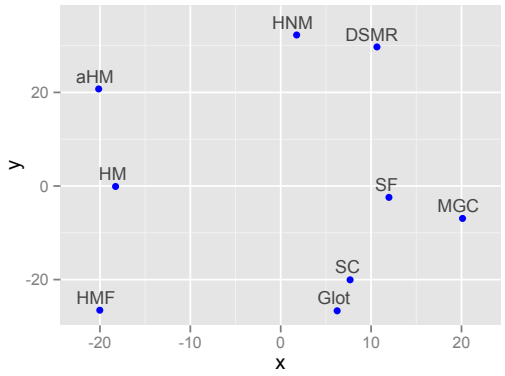
[11] S. Shechtman, and A. Sorin. "Sinusoidal model parameterization for HMM-based TTS system." In Proc. Interspeech, pp.805-808., Makuhari, Japan, September 2010.

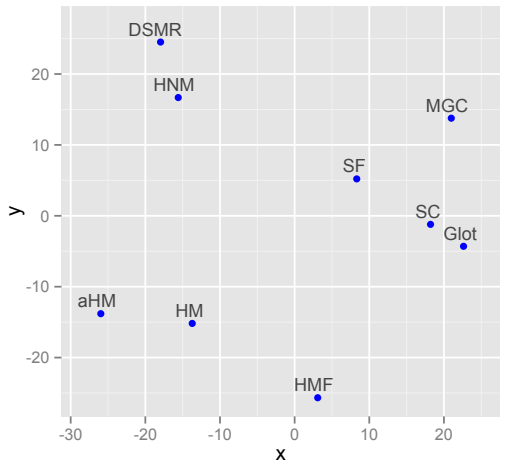
[12] Quatieri, F. T., "Discrete time speech signal processing", Pearson education, 427-439, 2004.

[13] P. Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics 20 (1987): 53-65.

[14] C. Mayo, R. A. J. Clark, and S. King."Multidimensional scaling of listener responses to synthetic speech." In Proc. Interspeech 2005, Lisbon, Portugal, September 2005.







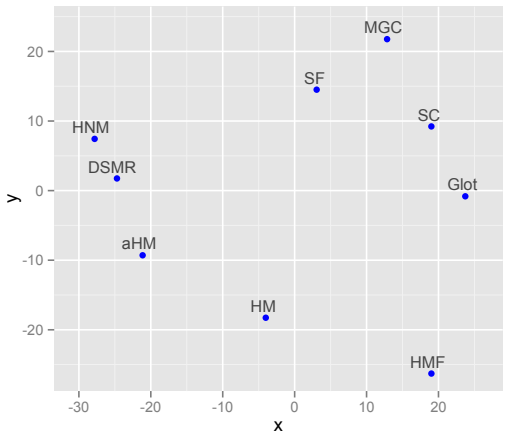
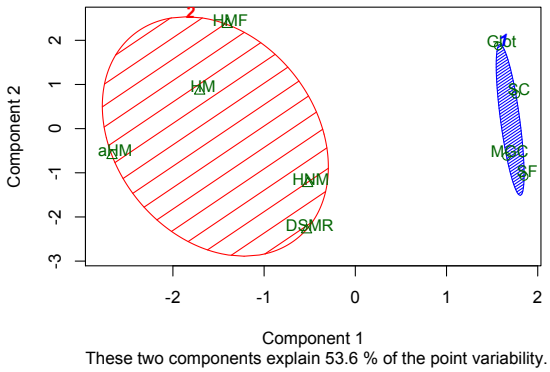
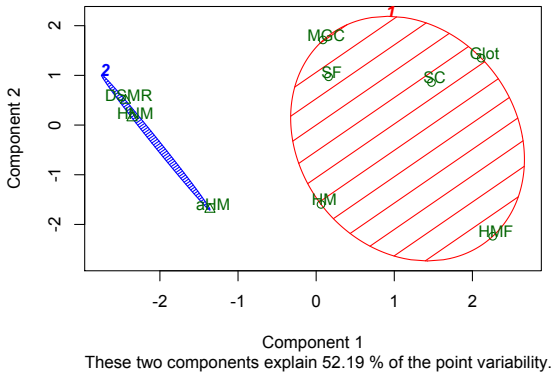
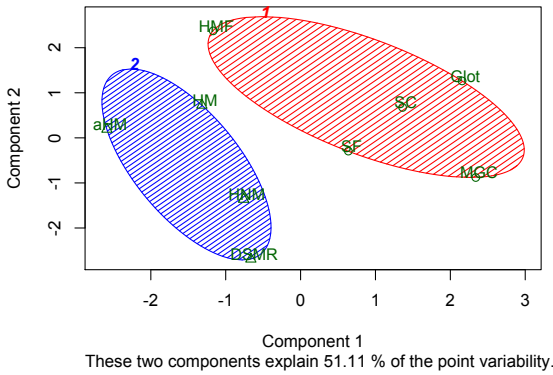


Figure 3: *MDS results for each section(up to down 1,2,3,4)*







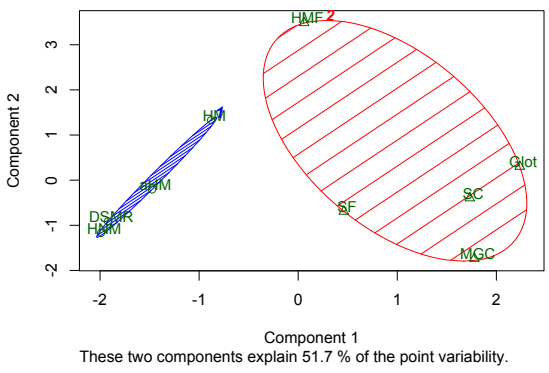


Figure 4: *K-means clustering results for each section (up to down 1,2,3,4)*