SPECTRAL DISTANCE COSTS FOR MULTILINGUAL UNIT SELECTION IN SPEECH SYNTHESIS

Hamurabi Gamboa Rosales, Oliver Jokisch and Rüdiger Hoffmann
Dresden University of Technology
Laboratory of Acoustics and Speech Communication, D-01062 Dresden, Germany
Hamurabi.Gamboa@ias.et.tu-dresden.de

ABSTRACT

The unit-selection module in concatenative TTS systems plays an important role regarding corpus synthesis. It has the main goal to minimize a composition of target and concatenation costs for a given phrase. We measured the concatenation cost through the spectral discontinuity perceptions, which are based on the spectral properties measures like: Linear spectral frequencies (LSFs), Multiple centroid analysis (MCA) and Mel-frequency cepstral coefficients (MFCCs). To determinate and evaluate the relationship between our spectral distance measures and distortion human perception, we report a perceptual experiment's guide to measure the correlation between human mismatch perception and spectral distance measures of concatenation costs in the multilingual, concatenative TTS system Papageno while testing the method for English, German and Spanish.

1. INTRODUCTION

In speech synthesis based on unit-selection, speech synthesis is produced by concatenating speech or acoustic units selected from a large-scaled database [1]. This database should be designed to cover as much as possible the phonetic and prosodic characteristics of a determined language or many languages. Consequently it increases the necessity to create and develop an efficient unit-selection process, which allows the multilingual TTS systems to handle a high volume data, variety and diversity of its content. At the moment that the TTS system synthesizes an utterance the unit-selection process tries to choose the best unit sequence from the database. Therefore the process to choose the best unit sequence can be done by an optimal unit-selection process and it is based on two costs: target cost and concatenation cost [2]. Both costs can be determined as the weighted sum of sub-costs, such as energy, pitch and duration for target sub-costs and LSF, MCAs and MFCCs for concatenation sub-costs. The research is a guide to evaluate and determinate which spectral measure from the concatenation sub-costs has the highest weight on multilingual unit selection process, to find the mismatch between two speech units in a given sentence in English, German or Spanish.

2. COST FUNCTIONS AND UNIT SELECTION

Hunt and Black [3] have described unit-selection as a search for a low cost candidate unit sequence. Therefore many investigators have proposed different cost functions.

2.1 Cost functions

Unit-selection process is divided into two cost functions. The first function is the target cost T^t, which is defined as an estimation of the mismatch between a recorded acoustic unit u_x and the predicted specification t_x [4] and it is calculated as the weighted sum of characteristic distances between the components of the target and candidate feature vector.

$$T^t(t_x, u_x) = \sum_{y=1}^{n} w_y^t T_y^t(t_x, u_x) \qquad (1)$$

Where n is the number of weighted target sub-costs, such as energy, pitch and duration, etc. The second function is the concatenation cost J^c, which is defined as an estimate of quality of the acoustic mismatch of the join between consecutive units (u_{x-1}, u_x) and it is calculated as the weighted sum of m concatenation sub-costs such as LSFs, MCAs and MFCCs, etc [4].

$$J^c(u_{x-1}, u_x) = \sum_{y=1}^{m} w_y^c J_y^c(u_{x-1}, u_x) \qquad (2)$$

In this paper we pay special attention on the concatenation cost, which is composed by the concatenation sub-costs LSFs, MCAs and MFCCs in a multilingual TTS system.

2.2 Unit Selection

The two cost functions previously defined compose a unit selection process. By the sum of the target cost and concatenation cost we can now determine the total cost function *C* for a sequence of *k* speech units:

$$C(t_x^n,u_x^n)=\sum_{x=1}^{k}\sum_{y=1}^{n}w_y^t T_y^t(t_x,u_x)+\sum_{x=2}^{k}\sum_{y=1}^{m}w_y^c J_y^c(u_{x-1},u_x) \quad (3)$$

The task of unit-selection process is to find the best sequence of *k* units so that it gives the minimum total cost of a given utterance.

3. SPECTRAL DISTANCE MEASURES

As we explained before, the concatenation sub-costs are determined by LSFs, MCAs and MFCCs spectral distance measures. Then we computed the concatenation sub-costs by a simple single frame (10 ms) distance, using only the final frame of the first unit and the initial frame of the second unit, because at this point is presented the join distortion.

3.1 Linear Spectral Frequencies (LSFs)

The LSFs are a parametric representation of all-pole spectrum (Vocal tract) and it was introduced as an alternative LPC spectral representation. The LSFs are defined as the root of the following two polynomials based upon the inverse filter *A(z)* [5].

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1}) \quad (4)$$

$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1}) \quad (5)$$

3.2 Multiple Centroid Analysis (MCAs)

MCAs are an alternative to formant estimation approach, which was proposed by Crowe [6]. We used MCAs with three centroids, these could be obtained by the expansion of the MCAs general formula:

$$e(k) = \sum_{w=a}^{b}(w-k)^2 S(w) \quad (6)$$

Where *S(w)* is the power spectrum signal , *k* is the centroid and *a,b* are the movil boundaries to obtain the minimum square error *e(k)* and of that form to obtain the centroid.

3.3 Mel frequency cepstral coefficient*s* (MFCCs)

FFT-based MFCC is also a popular acoustic parameter used in speech recognition and analysis. This is based on Mel-scale and tries to imitate the human listening processing. Mel-scale has the characteristic that it is linear to 1 kHz and logarithmic afterwards.

3.4 Parametric distance

The Euclidean distance *E(X,Y)* is defined as the straight line distance between two points or feature vectors as *X* and *Y*.

$$E(X,Y) = \sqrt{\sum_{i=1}^{m}(X_i - Y_i)^2} \quad (7)$$

4. STIMULI LISTENING TEST

Syrdal mentioned that a reliably higher discontinuity-detection rate for diphthongs is observed than for monophthong vowels [7]. Hence we tried to rate the concatenation mismatch by producing the diphthongs of each language. This was achieved by the concatenation of two diphonemes and the construction of one normal sentence per diphthong in the corresponding language. Then we produced different synthetic versions by varying one or two diphonemes per sentence, forming the diphthong and keeping the other diphonemes of the sentence. The join process of diphonemes was made for each language and the corresponding sentences pro diphthong were synthesized. For instance we synthesized five diphthongs for American English language like: ey(eI), ow(oU), ay(aI), aw(aU) and oy(OI), three for German language: ei(aI), eu(oI) and au(aU) and thirteen for the Spanish language: ai(aj), au(aw), ei(ej), eu(ew), oi(oj), ia(ja), ua(wa) , ie(je), ue(we), io(jo), uo(wo), iu(ju) and ui(wi). The diphthong ou was omitted because it is only used in an uncommon foreign word. Below the diphthongs and sentences for each language in their corresponding table are showed. In table 4.1,4.2 and 4.3 those words are highlighted, which do contain the corresponding diphthong.

Diphthong	Sentences
/ey/	She **take**s the bus.
/ow/	The **boat** has damage.
/ay/	He **ride**s a horse.
/aw/	It is so **loud**.
/oy/	I hear a **nois**e.

Table 4.1: English diphthongs and sentences

Diphthong	Sentences
/aI/	Das ist d**ei**n Auto. It is your car.
/OI/	Die K**eu**le ist kapput. The mallet is broken.
/aU/	Der R**au**m ist gross. The room is big.

Table 4.2: German diphthongs and sentences

Diphthong	Sentences	
/ej/	V**ei**ntidós euros.	Twenty-two euros.
/ja/	El p**ia**no hermoso.	The beautiful piano.
/we/	El j**ue**go de fútbol.	The soccer game.

Table 4.3: Spanish diphthongs and sentences

Because there are many diphthongs in Spanish, we mention only some sentences to show the diphthongs that were analyzed.

4.1 Test proceedings

There were around 10 listeners for the English listening test and they their age varied from 20 to 30. Some of them were native English speakers and the other had a good English proficiency. For the German listening test there were around 11 listeners, most of them students from the Dresden University of Technology and all of them were native German speakers from Saxony, also with an age between twenty and thirty. For the Spanish listening test there were around 10 listeners, all of them were native Spanish speakers and they also varied between twenty and thirty. The tests for all languages were composed by blocks of 30 stimuli sentences, which were previously selected from a database of stimuli sentences to get the best 30 stimuli sentences, also five sentences were duplicated to validate the listener scores. One block was created for each sentence in the corresponding language; see Tables 4.1, 4.2 and 4.3. Then the listener was played each test stimuli and we asked them to rate the quality join of each sentence on a scale of 1(worst) to 5(best). To support the task the listener were able to see the written sentence and the word which contained the join at every moment. Also as reference they could listen to the best join stimuli as 5 and the worst join stimuli as 1, which were previously selected by us from the test stimuli. Also they were allowed to listen to all the sentences as many times as they liked. Because the listening tests were a hard task to achieve in only one session, the listeners could complete the test in a few sessions.

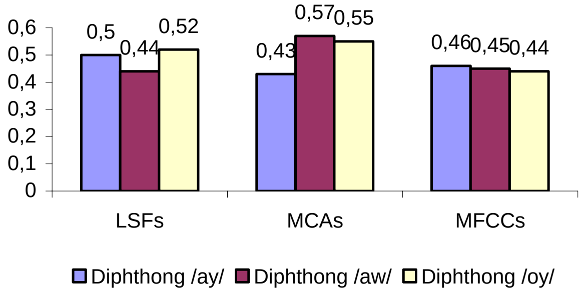
5. RESULTS AND CONCLUSION

The main goal is to determinate which concatenation sub-cost has the highest weight by the correlation of the human mismatch perception and the spectral distance measures. Below is showed the correlation function *r* [8] that we used to find it. The correlation was computed between the mean listener scores of each synthesized sentence version in the different languages and the spectral distance measures.

$$r = \frac{\sum_{i=1}^{n} y_i (x_i - x_m)}{\left[\sum_{i=1}^{n} (x_i - x_m)^2 \sum_{i=1}^{n} (y_i - y_m)^2 \right]^{\frac{1}{2}}} \qquad (8)$$

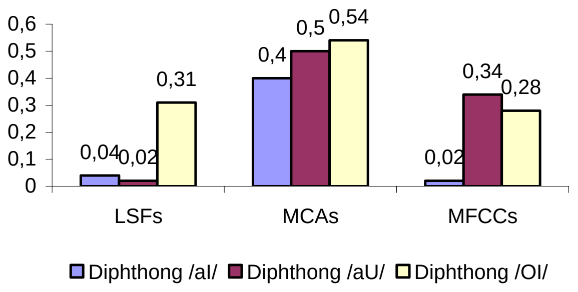
Where *n* are the number of spectral measures, x_i are the mean listener scores (MLS) per sentence, $\mathrm{x_m}$ the mean of MLS, y_i the distance concatenation costs between two consecutive units and $\mathrm{y_m}$ the mean of the concatenation costs. In this study all the correlation coefficients between all the concatenation sub-costs for every language were calculated by the Euclidean distance. However such a big result table was obtained that it could confuse the readers while trying understanding the results. So we decided to show that results of three diphthongs per language that presented more similarities to each other. Subsequently we selected for English /ay/, /aw/ and /oy/, for German /aI/, /aU/ and /OI/, and for Spanish /ej/, /ja/ and /we/. Below the figures 5.1, 5.2 and 5.3 show the correlation coefficients results per language with its corresponding diphthong.

Fig 5.1: Correlation between English MLS and concatenation sub costs.



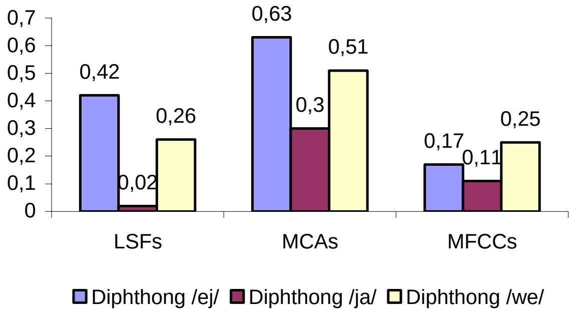
The figure 5.1 shows that the MCAs concatenation sub-cost for /aw/ and /oy/ diphthongs have higher correlation coefficients than LSFs and MFCCs. But also we notice that the correlation coefficients difference between MCAs, LSFs, and MFCCs is not great enough to make a conclusion.

Fig 5.2: Correlation between German MLS and concatenation distances.



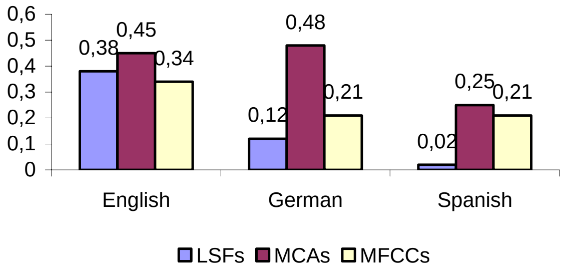
The MCAs concatenation sub-cost shows higher correlation coefficients than MFCCs and LSFs for the /aI/, /aU/ and /OI/ diphthongs. Also the MCAs correlation coefficients yield the 5% of significance for the three diphthongs, which demonstrates the high correlation between the mismatch human perception and the MCAs concatenation sub cost.

Fig 5.3: Correlation between Spanish MLS and concatenation distances.



Also MCAs concatenation sub-cost shows higher correlation coefficients than LSFs and MFCCs for the /ej/, /ja/ and /we/ diphthongs in Spanish and the MCAs correlation coefficients yield the 5% of significance for /ej/ and /we/ diphthongs. Below the figure 5.4 illustrates the mean of the correlation coefficients of each spectral measure in the corresponding language.

Fig 5.4: Correlation Coefficient Mean pro Language.



MCAs show the highest mean for every language, consequently we can conclude that MCAs has the highest correlation with the human mismatch perception in most of the cases. Hence the MCA must have the highest weight in the multilingual unit selection process. Afterward LSFs is determined as the second sub-cost with the highest weight in the multilingual unit selection process and MFCCs as the third concatenation sub-cost. Although MFCCs was found the worst concatenation sub-cost, we consider important the level of correlation that it showed in some cases. For this reason it is important to take into account the presented weight of MFCCs.

The results illustrate that the correlation between the concatenation sub-costs and human mismatch perception does not have a language dependency, because the MCAs showed a regular highest weight for the three languages. Below on the table 5.1 is showed the ranking of the three concatenation sub-costs that were analyzed for the multilingual unit selection.

Rank	*Concatenation Sub-Cost*
1	MCAs
2	LSFs
3	MFCCs

Table 5.1: Ranking for the three concatenation sub-costs

6. FUTURE WORK

MCAs concatenation sub-cost showed a higher correlation coefficients tendency than LSFs and MFCCs for every language. Nevertheless the LSF and MFCCs concatenation sub-costs showed also in some cases acceptable correlation coefficients or inclusive higher correlation coefficients than MCAs in some cases. Therefore the future work should survey which weights correspond to every sub-cost in each language to achieve an optimal concatenation cost function.

7. REFERENCES

[1] J. Vepa, S. King, and P. Taylor, “New objective distance measures for spectral discontinuities in concatenative speech synthesis,” in ICSLP, Denver, USA, 2002.

[2] J. Vepa, S. King, and P. Taylor, “Objective distance measures for spectral discontinuities in concatenative speech synthesis,” in ICSLP, Denver, USA, 2002.

[3] Andrew J. Hunt and Alan W. Black. “Unit Selection in a concatenative speech synthesis system using a large speech database” ICASSP,1:373-376,1996

[4] M. Beutnagel, M. Mohri, M. Riley. “Rapid unit selection from a large speech corpus for concatenative speech synthesis” in Proc. Eurospeech ’99, Budapest, Hungary, Sep 1999, pp. 607-610

[5] Lawrence R. and Biing-Hwang J. “Fundamentals of speech recognition,” Prentice Hall 1993

[6] Crowe A. and MA Jack. (1987), “Globally optimizing formant tracker using generalized centroids”, Electronic Letters, Vol 23, No. 19, pp 1019-1020.

[7] A.K. Syrdal “Phonetic effects on listener detection of vowel concatenation” in Ptoc. Eurospeech, (Aalborg, Denmark), 2001

[8]W. Hines and C. Montgomery, “Probability and Static in Engineering and Management Science,3rd. ed., CECSA 2004”,