A phonetic-contrast motivated adaptation to control the degree-of-articulation on Italian HMM-based synthetic voices

Mauro Nicolao[1], *Fabio Tesser*[2], *Roger K. Moore*[1]

[1]Speech and Hearing Group, Dept. Computer Science, University of Sheffield, UK
[2]Institute of Cognitive Sciences and Technologies, National Research Council, Padova, Italy
m.nicolao@dcs.shef.ac.uk, fabio.tesser@pd.istc.cnr.it, r.k.moore@dcs.shef.ac.uk

Abstract

The effectiveness of phonetic-contrast motivated adaptation on HMM-based synthetic voices was previously tested on English successfully. The aim of this paper is to prove that such adaptation can be exported with minor changes to languages having different intrinsic characteristics. The Italian language was chosen because it has no obvious phonemic configuration towards which human speech tend when hypo-articulated such as the mid-central vowel (schwa) for English. Nonetheless, low-contrastive attractors were identified and a linear transformation was trained by contrasting each phone pronunciation with its nearest acoustic neighbour. Different degree of hyper and hypo articulated synthetic speech was then achieved by scaling such adaptation along the dimension identified by each contrastive pair. The Italian synthesiser outcome adapted with both the maximum and the minimum transformation strength was evaluated with two objective assessments: the analysis of some common acoustic correlates and the measurement of a intelligibility-in-noise index. For the latter, signals were mixed with different disturbances at various energy ratios and intelligibility was compared to the standard-TTS generated speech. The experimental results proved such transformation on the Italian voices to be as effective as those on the English one.

Index Terms: hypo/hyper-articulated speech synthesis, Italian HMM-based synthesis, intelligibility enhancement, speech adaptation, statistical parametric speech synthesis.

1. Introduction

The observation that human talkers modify their speech production according to the environmental condition was established more than a century ago by Lombard [1].

According to theories such as the Lindblom's H&H (hypo-hyper) theory of speech production [2], such adjustments are controlled by the need of maximising the effectiveness of the communication minimising the effort at the same time. Mainly, these are driven by constant monitoring their effectiveness.

Moore's PRESENCE [3] model was one of the first attempts to import human-inspired ideas into a computational model, which could be also included in a complex automatic speech communication system. Following this comprehensive model, a first reactive synthesiser that could react to environmental disturbances by enhancing the contrast between competing phones was proposed [4]. This work was mainly motivated by the lack of expressivity observed in standard text-to-speech systems. The few expressive synthetic speech synthesisers are tuned to specific needs and therefore not able to react to different environmental conditions dynamically.

A first complete Computational model for Hyper and Hypo articulated speech synthesis (C2H), which would monitor its output in order to maximise the intelligibility in noise, was proposed in [5] and its effectiveness was tested with an HMM-based English speech synthesiser.

Recently, other approaches, which focus mainly on sound quality modifications to maximise the speech audibility in noise, have tackled the problem from different angles: synthetic Lombard-speech generation by manipulating the glottal source signal [6]; feature optimisation to increase the intelligibility of the parametric generated speech [7]; hyper and hypo-articulated speech synthesis by interpolating between 'ad hoc' recorded corpora [8].

Aim of this paper is to apply the C2H test adaptation technique [5], which relied upon the reduction/expansion of the vowel space using the schwa [ə] vowel, with minor changes to Italian. This language differs from English because it has no such clear low-energy attractors for hypo-articulated speech in its phonetic inventory.

In the following sections, the details of the adaptation on two existing Italian HMM-based voices (a female and a male one) are presented. Particular emphasis is used to describe the training process and to motivate the low-contrastive attractor choice. The Italian synthesiser outcome adapted both with the maximum and the minimum transformation strength was evaluated with two objective analyses: the extraction of some common acoustic correlates and the measurement of an intelligibility-in-noise index.

2. The Italian language

The adaptation process used in the C2H experimental part [5] took advantage of some characteristics of the English language in which a vowel exists, [ə], which is widely recognised as the most common reduced phonetic configuration in hypo-articulated speech.

The question therefore has surged whether low-contrastive configurations could be also found in the languages, such as Italian, where low-energy phones cannot be explicitly found.

Italian is a seven-vowel language with some peculiar differences with respect to English: i) the absence of low-energy phonemes such as /ə/ and /h/ in its phonemic inventory; ii) vowel acoustic realisations mostly stand close to the border of the vocalic triangle (F1-F2 chart); iii) the high variability of stress position in the word, along with the contrastive use of it; iv) the contrastive use of consonant geminations.

Even though Italian language does not exhibit schwa in his vocalic system, it can be observed (as allophones of some unstressed vowels) in spontaneous speech, in some reduction phenomena or in some local dialects [9]. Thus, the Italian hypo-

articulated speech is also assumed to contain one (or more) low-contrastive configurations towards which vowels are reduced. The main difference with English would be the selected target phones to train the linear transformation. The contrastive use of stressed/unstressed vowels and the consonant gemination, which mostly affects the phone duration, can be also controlled to reduce the acoustic distance between close phones.

As in English, formant shifting, spectral energy redistribution, speaking rate changes, pitch modification are the most common phenomena observed in Italian hyper/hypo-articulated speech.

3. The phonetic-contrast transformation

The basic principle that drives this transformation is that low-contrastive configurations exist for both human and synthetic speech. In such configurations, produced speech is less contrastive (i.e. phones merged together) and it becomes therefore less intelligible. On the contrary, when speech production moves away from these configurations, speech becomes clearer and intelligibility increases.

The adaptation process in C2H focused on low-level signal modifications, but it was also aware of the phonetic content of speech production along with the most likely competing phones for human understanding. Low-contrastive attractors were hypothesised for every phone to define the direction for hyper/hypo-articulated speech transformation. These attractors should be the most likely acoustic realisations towards which speech production converges when the effort has to be minimised (hypo articulated speech) and from which it moves when the intelligibility has to be maximised (hyper articulated speech). When phones are reduced to these low-contrastive configurations, the acoustic distance between the most likely competing phones is minimum. An interpolation/extrapolation along the key dimension of hypo/hyper-articulation could be obtained by controlling the distance from such attractors. The proposed adaptation was achieved with a linear transformation which also allows for a continuous adaptation.

Differently from what tested in the C2H experiments on English, no distinction was made between the adaptation on vowels and the one on consonants. In both cases, it consisted on identifying a phonetically relevant competitor for each phone and assuming that the Low-Contrastive (LC) configuration is reached by applying the half-strength transformation towards it. Ideally, this technique would map both competitors into the same acoustic realisation. The High-Contrastive (HC) configuration would be achieved by moving the operational point along the same direction but opposite strength.

A small difference between vowel and consonant adaptation stood in the criterion by which the competitors were chosen. An example for vowels is displayed in Figure 1.

3.1. Implementation in an HMM-based TTS system

In order to test the proposed hyper/hypo articulated speech adaptation, the adaptation was implemented to be applied to an HMM-based speech synthesiser. Figure 2 shows the functional diagram of the procedure used to create the target corpus and to train the transformation parameters.

Starting from the set of *Full context Labels (L0)* used to build the HMM-based voice, an all-contrastive version (*L1*) was obtained through the phonetic transformation. These labels were used to generate the acoustic features (*P1*) representing the low-contrastive acoustic space. The most likely models for

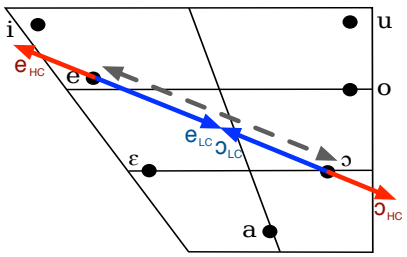


Figure 1: *Example of adaptations for [e] and [ɔ]. The blue lines refer to the transformations towards the Low-Contrastive (LC) point (hypo-articulated speech), the red lines to the ones towards the High-Contrastive (HC) (hyper-articulated). The dashed grey line shows the competitors used to train the transformation.*

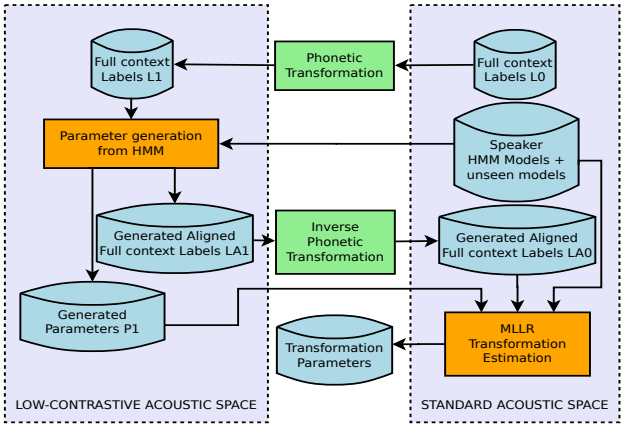


Figure 2: *Schematic diagram of the data preparation and the parameter transformation estimation.*

unseen-context phones were selected from the standard HMM models with a decision-tree clustering. The time-aligned version of *L1* (*LA1*) was mapped back into the standard phonetic domain (*LA0*). These labels along with the target *Generated Parameters (P1)* were used as reference to compute the *Transformation Parameters* with the Maximum Likelihood Linear Regression (MLLR) transformation estimation.

Once the parameters were obtained for the contrastive transformation, this would be applied with reduced strength (usually 50%) to the original HMM models to reduce standard synthesised speech to low-contrastive (hypo-articulated). Inverting the direction of the transformation as per [5], a high-contrastive (hyper-articulated) synthetic speech would be generated.

4. Italian TTS modules and voices

The base voice models used in the experiments were built using a modified version of the TTS software MaryTTS 5.0 [10]. The Italian modules and voices for MaryTTS [11] have been recently made available by the ISTC-CNR research institute and they comprehend: i) Italian lexicon and letter-to-sound rules, ii) context dependent part-of-speech tagger, iii) ToBI rules to predict symbolic prosody from text, iv) a customised version of the Italian Sampa phoneset [12].

The adaptation experiments were tested on two voice models both trained with phonetically and prosodic balanced speech corpora: *Lucia* (∼1400 sentences, ∼2 hours) recorded by a female speaker in a quasi-soundproof booth; *Roberto* (a commercial speech corpus available for research purposes, ∼1900 sentences, ∼3 hours) recorded by a professional male speaker.

Both voice models adopted MaryTTS as linguistic front-end for extracting monophone and full context labels. The HTS toolkit version 2.2 [13] was used to model the signal spectrum with Mel-Generalised Cepstral features, the fundamental frequency with multi-space probability distribution (MSD) [14], and the voicing strengths for mixed excitation [15] with continuous probability distribution.

The voices were trained with: i) default speaker-dependent HTS parameters, ii) decision tree based state clustering, iii) separate streams to model each of the static, delta and delta-delta features, iv) single Gaussian models.

Both voices are proven to have high-quality characteristics. *Lucia* is actively employed in robot-human interactions within the EU-funded project ALIZ-E[1], while the commercial voice *Roberto* has received good scores from some initial informal listening test.

5. Experiments

First a set of competing phone pairs was selected. For each vowel, the competitor was the one in opposite position across the vowel triangle (see Figure 1). On the other hand, the consonant competitor pairs were motivated by a study about perceptual confusion/discrimination on Italian consonants in noise [16]. Similarly to [17], Caldognetto listed the Italian consonants that are more likely to be mistaken in several noisy conditions. Following these guidelines, the most confusable consonant pairs were chosen to be the competitor pairs. Further contrastive pairs were motivated by the contrastive use of consonant gemination in Italian. The geminated consonants were therefore mapped into the corresponding non-geminated ones.

A summary of the substitutions is displayed in Table 1.

Table 1: *Vowel and consonant mapping. STD column contains the original phones and CTR has the contrastive ones. Geminate consonants are mapped to the corresponding non-geminate ones.*

STD	CTR	STD	CTR	STD	CTR
[a]	[u]	[f]	[p]	[dz]	[dʒ]
[e]	[o]	[t]	[k]	[dʒ]	[dz]
[i]	[ɔ]	[k]	[t]	[g]	[dʒ]
[o]	[e]	[ts]	[s]	[z]	[g]
[u]	[a]	[s]	[ts]	[l]	[ʎ]
[ɛ]	[o]	[tʃ]	[s]	[ʎ]	[l]
[ɔ]	[e]	[ʃ]	[tʃ]	[m]	[n]
[j]	[ɔ]	[b]	[d]	[n]	[m]
[w]	[a]	[d]	[b]	[ɟ]	[m]
[p]	[f]	[v]	[b]	[r]	[m]

After that, the relative adaptation for the *Lucia* and *Roberto* voices was obtained as per the method described in Section 3.1.

The same framework used to test the C2H model was also applied to evaluate this adaptation. A common test set of 200 text sentences (not included in the training set) were used to generate the test utterances which was then assessed with objective evaluations.

As mentioned above, the transformation has to be applied to the standard TTS voice (STD) with the appropriate strength in order to reach the correct low-contrastive and high-contrastive operational points. The generated speech signals were synthesized with three degree of articulation: a) HYO the lowest-contrastive configuration which is still intelligible in clean environment (60% of the total strength); b) STD the standard TTS outcome (no adaptation); c) HYP the highest-contrastive configuration which is not affected by too sever artefacts (60% of the total inverse strength).

Before the evaluation, all signals were normalised to have a constant RMS (-24 dBFS).

Thereafter, the three different kinds of speech signals were analysed using automatic tools to extract acoustic correlates and the results compared with what observed in literature for hyper and hypo-articulated speech. These speech signals were also mixed with three different disturbances: i) a real car noise, ii) a babble noise recorded in a large-size room, iii) 2-3 competing different language (English) talkers. In order to normalise the speech energy with respect to the noise, the Segmental Signal to Noise Ratio (SSNR) [18] was computed and the disturbance was amplified to have constant mean SSNR. Three SSNR levels has been taken into consideration for these experiments: 1 dB, -4 dB, and -9 dB.

6. Results

In order to evaluate the proposed adaptation performance, two types of analysis on the three types of synthesiser outcome are provided: an acoustic and an intelligibility one.

6.1. Acoustic analysis

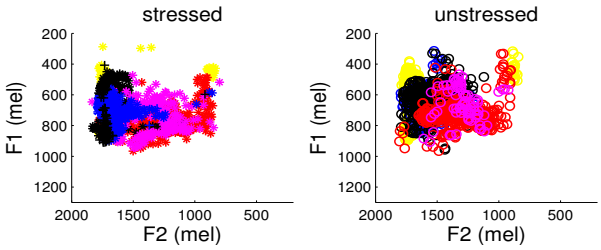
The first assessment of the adaptation effects on the acoustic signal was done by plotting the first two vowel formants (F1 and F2) of synthesiser outcome with the three degrees of articulations (HYP, HYO, and STD) (Figure 3 and Figure 4).

It is clear from these plots how both TTS voices move from the confused and centralised positions of the HYO configuration (Figure 3a and 4a) to the more separate and recognisable ones of HYP (Figure 3a and 4a). The stressed HYO vowels in *Roberto* aggregate in three main positions rather than an unique central one. This confirms the idea that the low-contrastive configuration is not an unique position close to [ə], but it is an intermediate position depending on the surrounding phones. All these behaviours emerged spontaneously from the adaptations without any assumption in the training but the control of phonetic contrast. Transformations seem to achieve a more effective reduction/expansion on the *Roberto* voice. It is worth to notice that the *Roberto* vowel variance (Figure 4) is quite limited with respect to the *Lucia* one (Figure 3). Indeed, the former was created using a professional speaker's voice whereas the latter denotes some regional accent influence. Moreover, the amount of recorded corpus is different: *Roberto* training corpus was one-third bigger the *Lucia* one.

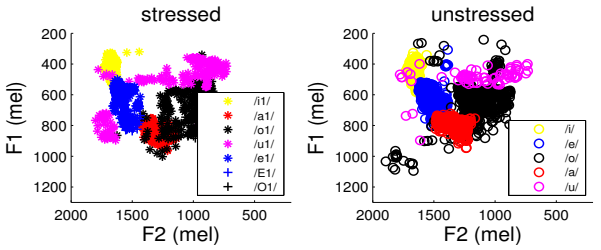
Even though two adaptations contribute to modify the signal at the same time, the vowel charts behave similarly to what observed for the schwa-based ones in [5].

Another type of acoustic analysis was performed by extracting some acoustic parameters, which are proved to be correlated to the degree of articulation of speech [19, 20], from the audio generated by the implemented adaptations:

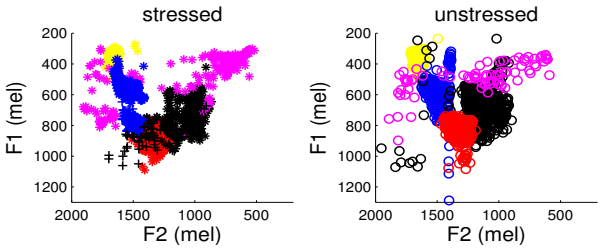
[1] http://www.aliz-e.org/



(a) HYO

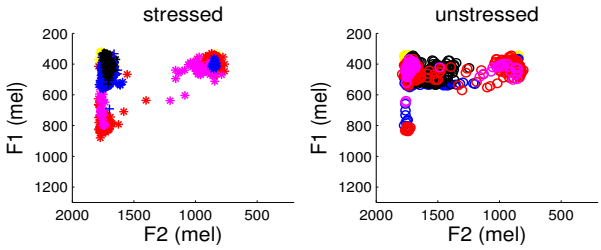


(b) STD

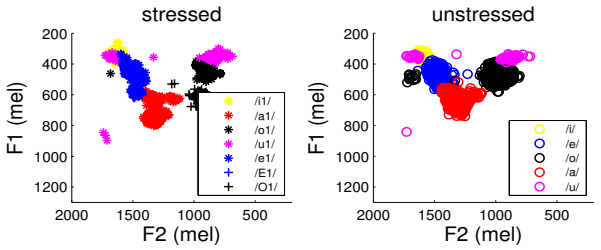


(c) HYP

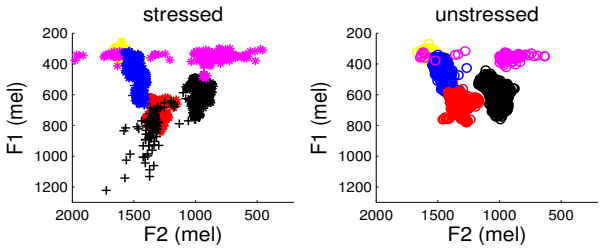
Figure 3: Effect of the *HYO* (Figure 3a) and the *HYP* (Figure 3c) adaptation applied to the *Lucia* voice. The plot for STD voice vowels is also plotted (Figure 3b) for reference. SAMPA symbols are used in the legend.



(a) HYO



(b) STD



(c) HYP

Figure 4: Effect of the *HYO* (Figure 4a) and the *HYP* (Figure 4c) adaptation applied to the *Roberto* voice. The plot for STD voice vowels is also plotted (Figure 4b) for reference. SAMPA symbols are used in the legend.

duration parameters : the Mean Sentence Duration, *MSD*, and the Mean Phone Duration (without pauses), *MPD*, which normally increase in human speech with the degree of articulation;

spectral parameters: the Long Term Average Spectrum, *LTAS13*, the spectral tilt, *Sp.Tilt*, the spectrum Centre of Gravity *Sp.CoG*, and the vowel space area (*F1F2 area*) which usually show a shifting towards high frequency;

pitch parameters: the average fundamental frequency value, *F0*, and its range, *F0 range* which should both increase accordingly to the degree of articulation.

The average result values are shown in Table 2 for *Lucia* and in Table 3 for *Roberto*. In both tables, the clearest modifications are observed in the vowel space (*F1F2 area*) expansion/reduction. Even though this was imposed by design, nonetheless this observation, together with Figure 3 and Figure 4, it proves the adaptation behaves correctly.

Other evident differences between the three sets of audio files appear in the spectrum energy shift (e.g. *Sp.CoG* and *Sp.Tilt*) and in the duration (*MSD* and *MPD*). The latter proves the tendency of the automatic system to elongate the speech production to increase phonetic contrast and vice-versa.

In conclusion, from the acoustic analysis, it can be affirmed that the proposed transforms are behaving according to what observed in human speech production.

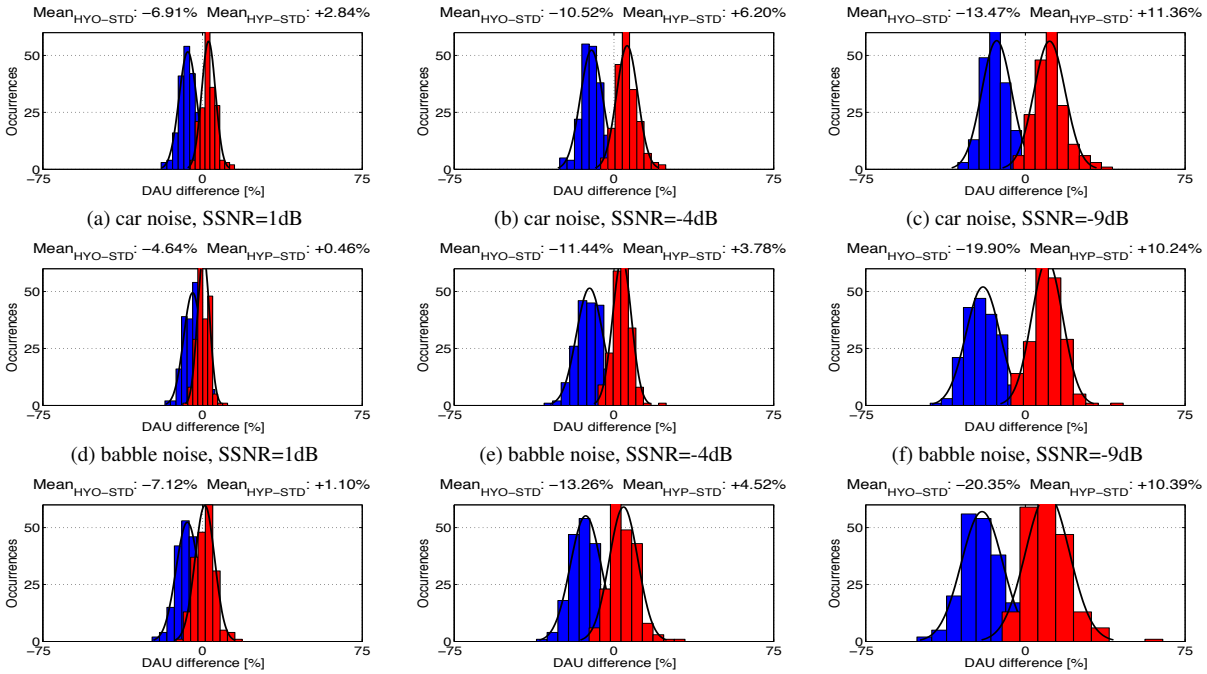
Table 2: Acoustic analysis of the three degrees of adaptation on the *Lucia* voice. In parenthesis the difference with STD.

Type of analysis	**HYO**	**STD**	**HYP**
MSD [s]	5.75 (-5.1%)	6.06	6.38 (+5.3%)
MPD [s]	0.078 (-2.5%)	0.08	0.083 (+3.7%)
LTAS13 [dB SPL]	47.7 (-9.3%)	52.6	58.3 (+10.8%)
Sp.Tilt [dB/dec]	-5.6 (+7.7%)	-5.2	-4.7 (-9.6%)
Sp.CoG [Hz]	394.1 (-27.9%)	546.2	835.9 (+53.0%)
F1F2 area [Hz^2]	14115 (-90.1%)	142401	203959 (+43.2%)
F0 [Hz]	197.3 (-3.4%)	204.3	210.3 (+2.9%)
F0 range [Hz]	138-225 (-23.6%)	133-247	134-276 (+24.8%)

6.2. Intelligibility evaluation

Intelligibility is strongly correlated with the effort involved in human speech production: the more adverse the condition, the higher the degree of articulation. Therefore, the proposed adaptation should also control the intelligibility in noisy conditions.

The HYO,HYP, and STD test files, mixed with disturbances as explained in Section 5 were processed with the Dau method [21] to assess speech intelligibility (DAU). Automatic intelligibility-assessment methods are quite important to score the speech synthesis quality. Even though most of them mainly measure the audibility of a signal without taking into account of

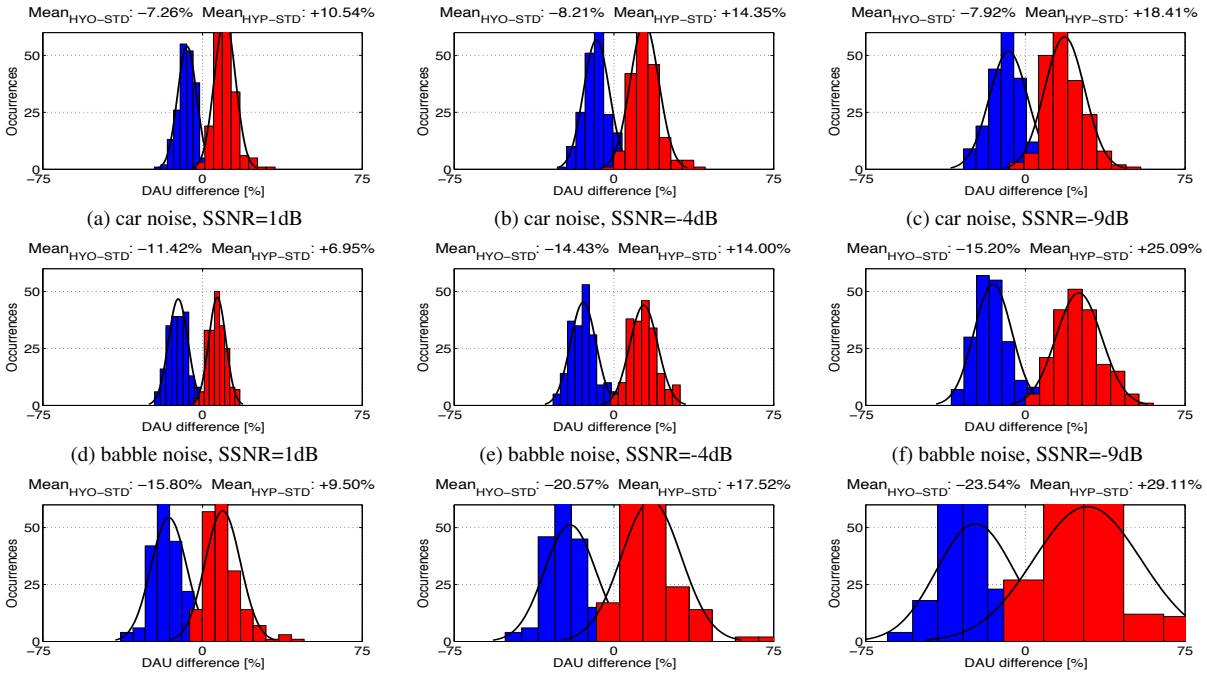


(a) car noise, SSNR=1dB (b) car noise, SSNR=-4dB (c) car noise, SSNR=-9dB

(d) babble noise, SSNR=1dB (e) babble noise, SSNR=-4dB (f) babble noise, SSNR=-9dB

(g) competing talker, SSNR=1dB (h) competing talker, SSNR=-4dB (i) competing talker, SSNR=-9dB

Figure 5: Objective evaluation of the Hyper (*HYP*) and Hypo (*HYO*) adaptation applied to the *Lucia* voice: distribution of the DAU differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms). Every row shows results for a different kind of noise and every column is related to the same SSNR.



(a) car noise, SSNR=1dB (b) car noise, SSNR=-4dB (c) car noise, SSNR=-9dB

(d) babble noise, SSNR=1dB (e) babble noise, SSNR=-4dB (f) babble noise, SSNR=-9dB

(g) competing talker, SSNR=1dB (h) competing talker, SSNR=-4dB (i) competing talker, SSNR=-9dB

Figure 6: Objective evaluation of the Hyper (*HYP*) and Hypo (*HYO*) adaptation applied to the *Roberto* voice: distribution of the DAU differences (in percentage) between HYO and STD speech (blue histograms), and between HYP and STD speech (red histograms). Every row shows results for a different kind of noise and every column is related to the same SSNR.

Table 3: Acoustic analysis of the three degrees of adaptation on the *Roberto* voice. In parenthesis the difference with STD.

Type of analysis	**HYO**	**STD**	**HYP**
MSD [s]	4.72 (-14.2%)	5.50	6.28 (+14.2%)
MPD [s]	0.06 (-16.7.5%)	0.072	0.083 (+15.3%)
LTAS13 [dB SPL]	44.7 (-6.9%)	48.0	56.3 (+17.3%)
Sp.Tilt [dB/dec]	-6.3 (+8.6%)	-5.8	-4.9 (-15.5%)
Sp.CoG [Hz]	434.5 (-30.6%)	625.8	947.0 (+51.33%)
F1F2 area [Hz^2]	469 (-99.6%)	124518	143156 (+15.0%)
F0 [Hz]	119.7 (+2.9%)	116.3	112.7 (-3.1%)
F0 range [Hz]	73-143 (-6.7%)	68-143	67-162 (+26.6%)

actual phonetic content of the signal, Dau's index is proved to be quite correlated to human understanding performances [22].

The DAU differences (in percentage) between HYO/HYP and STD are plotted in Figure 5 and 6. The figures show a clear intelligibility improvement/reduction with respect to the standard TTS voice in all types of noises for both voices. Signal modifications are more significant at medium/high levels of noise and when applied to the *Roberto* voice. On average, the intelligibility deviation from the STD voice is around 10% for both the HYO and the HYP adaptations.

7. Conclusions

In this paper, the same adaptation technique of the C2H model was applied to Italian to confirm its validity on a language with different phonological characteristics from English. A phonetic-contrast motivated transformation from standard to low-contrastive phone realisations was proposed on the basis of the vowel positions on the vocalic triangle and the perceptual discrimination of consonant pairs for this language.

The objective analyses confirm that the estimated transformation parameters can model different degrees of articulation for speech from low-contrastive (hypo-) to high-contrastive (hyper-articulated). Acoustic analyses on the generated speech confirm that the deviation of some relevant acoustic speech cues from the standard articulated speech, like vowel space expansion, frequency distribution, spectral tilt, speaking rate and pitch, follows the literature results when the hyper/hypo articulated models were used instead of the standard models. Moreover, intelligibility tests in noise condition have shown the increasing of the intelligibility of the generated hyper articulated speech with respect to the speech generated with the standard HMM models.

Hence, results indicate that the transformation, when applied to the two Italian HMM-based voices, is indeed effective, even if Italian does not have clear low-energy attractors for hypo-articulated speech in its phonetic inventory, such as schwa the for English.

8. Acknowledgements

The research leading to these results was funded by the EU-FP7 network SCALE (ITN-213850) and by EU-FP7 project ALIZ-E (ICT-248116). The authors would like to thank MiVoQ for supporting this research and providing the *Roberto* voice.

9. References

[1] É. Lombard, "Le Signe de l'Elevation de la Voix - The sign of the rise in the voice," *Ann. Maladiers Oreille, Larynx, Nez, Pharynx - Annals of diseases of the ear, larynx, nose and pharynx*, vol. 37, pp. 101–119, 1911.

[2] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory," *Speech production and speech modelling*, vol. 55, pp. 403–439, 1990.

[3] R. K. Moore, "PRESENCE: A Human-Inspired Architecture for Speech-Based Human-Machine Interaction," *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1176–1188, Sep. 2007.

[4] R. K. Moore and M. Nicolao, "Reactive Speech Synthesis: Actively Managing Phonetic Contrast Along an H&H Continuum," in *ICPhS 2011*, Hong Kong, China, Aug. 2011, pp. 1422–1425.

[5] M. Nicolao, J. Latorre, and R. K. Moore, "C2H: A Computational Model of H&H-based Phonetic Contrast in Synthetic Speech," in *INTERSPEECH 2012*, 2012.

[6] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Analysis of HMM-Based Lombard Speech Synthesis," in *INTERSPEECH 2011*, Florence, Italy, Aug. 2011, pp. 2781–2784.

[7] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Mel Cepstral Coefficient Modification Based on the Glimpse Proportion Measure for Improving the Intelligibility of HMM-Generated Synthetic Speech in Noise," in *INTERSPEECH 2012*, Portland, OR, Jun. 2012, pp. 1–4.

[8] B. Picart, T. Drugman, and T. Dutoit, "Continuous control of the degree of articulation in HMM-based speech synthesis," in *INTERSPEECH 2011*, Florence, IT, 2011, pp. 1797–1800.

[9] F. A. Leoni, F. Cutugno, and R. Savy, "The vowel system of Italian connected speech," in *ICPhS 1995*, B. P. Elenius K., Ed., vol. 4, Stockholm, 1995, pp. 396–399.

[10] M. Schröder, M. Charfuelan, S. Pammi, and I. Steiner, "Open source voice creation toolkit for the MARY TTS Platform," in *INTERSPEECH 2011*, Florence, Italy, 2011.

[11] F. Tesser, G. Paci, G. Sommavilla, and P. Cosi, "A new language and a new voice for MARY-TTS," in *9th national congress, AISV (Associazione Italiana di Scienze della Voce)*, Venice, Italy, 2013.

[12] "SAMPA for Italian," 1989 (accessed May 21, 2013). [Online]. Available: http://www.phon.ucl.ac.uk/home/sampa/italian.htm

[13] 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," in *The 6th International Workshop on Speech Synthesis*. Citeseer, 2007, pp. 294–299.

[14] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *ICASSP 1999*. IEEE, 1999, pp. 229–232 vol.1.

[15] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed Excitation for HMM-based Speech Synthesis," in *Eurospeech*, 2001.

[16] E. M. Caldognetto, K. Vagges, and F. Ferrero, "Intelligibilità e confusione consonantiche in Italiano," *Rivista Italiana di Acustica*, 1988.

[17] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *JASA*, Jan. 1955.

[18] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Trans. Speech, Audio & Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[19] V. Hazan and R. E. Baker, "Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions," *JASA*, vol. 130, no. 4, p. 2139, 2011.

[20] R. J. J. H. van Son and L. C. W. Pols, "An acoustic description of consonant reduction," *Speech Communication*, vol. 28, no. 2, pp. 125–140, Jun. 1999.

[21] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. I. Model structure," *JASA*, vol. 99, no. 6, pp. 3615–3622, Jun. 1996.

[22] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Evaluation of Objective Measures for Intelligibility Prediction of Hmm-Based Synthetic Speech in Noise," in *ICASSP 2011*, Prague, May 2011.