COMPLEMENTARY CUES FOR SPEECH RECOGNITION

Alain Soquet, Marco Saerens, and Véronique Lecuit
Université Libre de Bruxelles, Belgium

ABSTRACT

In this paper, we discuss the role of different complementary cues at three different levels: acoustic, aerodynamic and articulatory on a speech recognition task. The recognition system is a standard speaker dependant HMM recognizer. The corpus is based on CVCV sequences uttered by one speaker. It is shown that aerodynamic and articulatory cues used in combination with Mel frequency cepstral coefficients lead to a substantial increase of the system recognition score.

1. INTRODUCTION

One possible way to continue improving the performances of statistical speech recognizers is to complement the traditional acoustic cues with phonological, articulatory or audio-visual cues. These cues are difficult to obtain directly. To avoid this problem, some make use of phonological models, such as Deng et al. [3], others use a representation in phonetic features, such as Koreman et al. [6]. Zlokarnik [11] demonstrates the usefulness of using articulatory cues derived from measurements of an electro-magnetic articulometer for speech recognition.

In previous work, Soquet and Saerens showed the benefit of adding articulatory information in the case of identification of place of articulation of plosives [8] and vowels [9]. The cues they used were extracted from the speech signal by using acoustic-to-articulatory inversion.

Since we now have access to a wide range of direct articulatory measurements, it is possible to study the role of different complementary cues measured simultaneously with the speech signal.

Our goal in the present study is to present a speech recognition experiment using a wide set of cues obtained by different measurement techniques and to assess their role in speech recognition.

The different measurements have been collected simultaneously during the recording of the corpus with (i) an electropalatograph, which gives the contact points of the tongue with the palate, (ii) aerodynamic measurement (intra-oral pressure), and (iii) an electromagnetic articulometer which gives us the jaw opening and lip movements.

The cues derived from these measurements give direct information on the behavior of articulators during speech production such as the contact profile for the tongue, the opening of the jaw, and the movements of the lips. The intra-oral pressure is given because it has proven to be of valuable interest for the understanding of speech production.

The experiment compares the recognition scores as a function of the cues available to the system. The result of this comparison indicates (i) which cues are of interest for speech recognition, and (ii) to which extend complementary cues can have an effect on identification scores.

2. MATERIAL

2.1. Measurements

The experimental setup consists of a Physiologia workstation (Teston and Galindo [10]), gathering simultaneously the acoustic, electropalatographic, aerodynamic, and electro-magnetic articulometric information.

2.1.1. Acoustic. The speech signal has been sampled at 16000 Hz. Due to experimental conditions, such as the background noise of the recording apparatus, the recordings do not achieve soundproof room quality.

2.1.2. Electropalatographic. We used the electropalatograph developed at the University of Reading, (Hardcastle [4]), which is constituted from an artificial palate fitted with 62 electrodes distributed in rows from the alveolar to the palatal-velar limit. For each electrode a binary signal describes whether the electrode is in contact with the tongue or not. The state of each electrode is sampled at 2000 Hz.

2.1.3. Aerodynamic. The intra-oral pressure was measured through a 2 mm diameter rubber catheter going through the nasal cavity, and hanging in the oro-pharynx (see Figure 1). This catheter was connected to a pressure sensor of the Physiologia workstation. The value of intra-oral pressure is sampled at 2000 Hz.

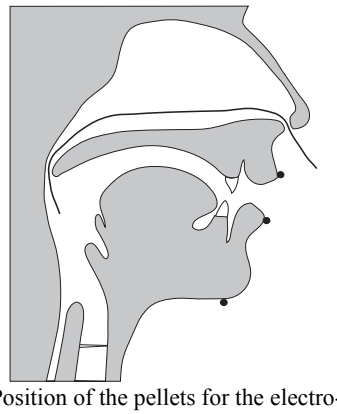

Figure 1. Position of the pellets for the electro-magnetic articulometer and position of the pressure catheter.

2.1.4. Electro-magnetic articulometry. The system used for the experiment is a Movetrack, developed by Branderud [1]. This system was used to collect information on the movement of three pellets: the upper and lower lips, and the jaw. The pellets were glued respectively on top of the upper lip, at the bottom of the lower lip, and on the chin (see Figure 1). The recorded signals correspond to the distances between the pellets and the two coils mounted on the helmet and are sampled at 2000 Hz. Both signals have been subsequently transformed into X and Y coordinates.

2.2. Corpus

The corpus is made of consonant-vowel-consonant-vowel sequences ($C_1V_1C_2V_2$) where C_1 and C_2 are chosen among [p, t, k, b, d, g, f, s, ʃ, v, z, ʒ, m, n, j, ʁ] and V_1 and V_2 among [i, e, a, o, u, y]. The total combination of these phonemes gives a total of 9216 sequences.

The objective was to record a corpus of acceptable length, with a minimum effect on phoneme distribution and covering a maximum of the variability of the complete combination. Therefore, we have defined 4 sets of 384 sequences in such a way that the couples C_1V_1, V_1C_2 and C_2V_2 are evenly distributed inside each set and between the different sets.

The resulting 1536 sequences have been recorded by one male native french speaker.

3. METHOD

3.1. Cues

All cues have been synchronously computed with an interval of 10 ms.

3.1.1. Acoustic (*MFCC*). 12 Mel frequency cepstral coefficients were computed on a 25.6 ms windowed speech signal [2].

3.1.2. Electopalatographic (*EPG*). In order to use the EPG data in a speech recognition system, the large number of electrode data has to be reduced to a few parameters. We choose the method presented by Recasens et al. [7]. Two cues are defined along the anterior-posterior dimension (the anteriority coefficient AC and the posteriority coefficient PC) and along the lateral-central dimension (the centrality coefficient CC). The value of these coefficients increases as the linguo-palatal contacts become more anterior (AC), posterior (PC) or central (CC). The definition of the coefficients was adapted to our palate.

The coefficients were computed at the moment corresponding to the middle of the window used for the acoustic analysis.

3.1.3. Aerodynamic (*Pio*). The value of the intra-oral pressure was downsampled from 2000 Hz to 100 Hz and used as a cue.

3.1.4. Electro-magnetic articulometry (*MVT*). The X and Y pellet positions were downsampled from 2000 Hz to 100 Hz and used as cues.

3.2. Hidden Markov Modeling

Each phoneme was modeled by a basic left-to-right hidden Markov model with three identical states. For each state, the emission probabilities are generated by a multi-dimensional Gaussian function with a diagonal covariance matrix. A distinction was made between the phonemes at the beginning of the sequence (C_1 and V_1) and the phonemes at the end of the sequence (C_2 and V_2), and between the initial and the final silence.

We used the Viterbi algorithm [5] both for the estimation of the model parameters and during the recognition phase. Notice that the training sets have not been pre-segmented.

Depending on the experiment, the observation vector was a concatenation of different cues.

3.3. Experiments

In order to study the effectiveness of these different representations we carried out nine experiments corresponding to nine different combinations of the different cues. Table 1 summarizes the cues used for each experimental setup. The first setup corresponds to an experiment where only articulatory cues are used to model speech, the last setup corresponds to an experiment where all the cues are used.

Table 1. Cues used for each experimental setup.

Setup	*MFCC*	*Pio*	*MVT*	*EPG*
1	0	1	1	1
2	1	0	0	0
3	1	0	0	1
4	1	0	1	0
5	1	0	1	1
6	1	1	0	0
7	1	1	0	1
8	1	1	1	0
9	1	1	1	1

For each experiment, we used a jack-knife approach. The corpus is composed of four balanced sets of CVCV sequences. Each set is successively used as test set and the remaining three as training sets.

The recognition task was word recognition, and the vocabulary consists in the 1536 sequences of the corpus.

4. RESULTS

4.1. Sequence recognition

Figure 1 displays the results for each setup, and for each test set. One can see that the results between sets are comparable as expected. Therefore, in the rest of the paper, the recognition scores are computed on all test sets together.

The sequence recognition results for the *MFCC* alone are below 50 percent. This low performance can be explained by the complexity of the recognition task: (i) all words of the

vocabulary have, by construction, the same structure, (ii) the size of the vocabulary for the recognition task is important (1536 word candidates), and (iii) the Hidden Markov Modeling is very basic. For the complementary cues alone the results are below 40 percent. This is not surprising, and corresponds to the fact that vowels are poorly modeled with the complementary cues alone. Nevertheless, this poor score is not too far from the score achieved when only *MFCC* cues are used.

As can be seen on Figure 2, the influence on the recognition score is positive for all complementary cues. The most important influence is systematically provided by the *EPG* cues with an increase of more than ten percent over the comparable experiment without the *EPG* cues. When used, the *Pio* also provide an increase in performances. The *MVT* alone with the *MFCC* is the only setup where the performance of the system does not improve. On the contrary, when it is added in conjunction with another complementary cue, the performance improves.

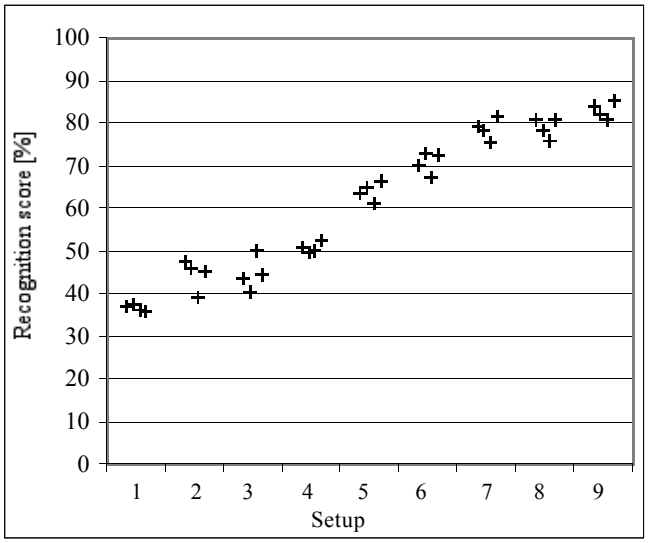

Figure 2. Sequence recognition scores as a function of the experimental setup.

4.2. Consonant and vowel recognition

The recognition score for consonants and vowels in first position (C_1 and V_1) and second position (C_2 and V_2) in the sequence are presented in Table 2.

The vowels in both positions are appropriately recognized in all setups but the first one which does not include the *MFCC* coefficients. This corresponds to the fact that vowels are poorly modeled by complementary cues alone.

Consonant recognition scores are highly improved when *EPG* cues are present. This is easily understandable as these cues are known to reflect the place of articulation of the consonants in the alveolar to the palatal-velar limit.

The *Pio* cue is mostly useful in the case of C_1, and *MVT* cues used alone are insufficient to help the system. Both these cues have a stronger influence on the recognition score when they are used in association with *EPG* cues.

Complementary cues are useful for consonants at variable degrees. Their positive influence on the recognition score is mostly visible in the case of C_1. This could be due to the case that C_1 being surrounded by a silence and a vowel, its boundaries are harder to identify than those of C_2. Complementary cues can help set these boundaries and lead to a better modeling.

Table 2. Consonant and vowel recognition scores [%] (recognition of the 1536 sequences with a jack-knife procedure).

Setup	*MFCC*	*EPG*	*Pio*	*MVT*	C_1	V_1	C_2	V_2	Total
1	0	1	1	1	55.8	75.4	65.1	83.9	36.8
2	1	0	0	0	56.8	94.4	79.2	95.4	44.6
3	1	0	0	1	53.1	92.2	80.3	94.5	44.7
4	1	0	1	0	62.7	94.8	79.8	96.9	50.7
5	1	0	1	1	76.2	95.0	82.7	96.5	64.0
6	1	1	0	0	81.4	92.6	87.3	94.2	70.8
7	1	1	0	1	87.9	94.9	88.8	95.5	78.7
8	1	1	1	0	89.1	94.0	89.2	95.2	79.0
9	1	1	1	1	91.8	95.5	90.4	96.4	83.0

4.3. Feature recognition

In order to understand better the behavior of the system as a function of the cues, we have computed the recognition scores for some important features like place and manner of articulation, and voicing. These features separate the different phonemes present in the corpus as described in table 3.

Table 3. Separation of the phonemes of the corpus with the three features follows (voiced consonants appear in bold characters).

		Place of articulation									
		Bilabial		Labio-dental		Alveolar		Post-alveolar		Velar	
Manner	Plosive	p	**b**			t	**d**			k	**g**
	Fricative			f	**v**	s	**z**	ʃ	**ʒ**		
	Nasal		**m**				**n**				
	Liquid						**l**				**ʀ**

Table 4 summarizes the results for each experiment.

As expected, *EPG* cues are effective to distinguish the place of articulation. *Pio* is also effective to a lesser extent. *MVT* cues are only effective when used in combination with *EPG* or *Pio*. If all the cues are combined together (setup 9), the system achieves a score of 94%.

Manner of articulation is well recognized with *EPG* and *Pio* in the case of C1. *MVT* cues are only useful in combination with *EPG* or *Pio*.

Voicing is best recognized when *Pio* and *MFCC* are used together. Surprisingly, in the case of setup 1, the system was able to distinct voicing with no acoustic cues. This can be explained by the fact that the value of the *Pio* can give useful information to discriminate between the voiced and voiceless consonants with the same place and manner of articulation.

Table 3. Feature recognition scores [%]
(recognition of the 1536 sequences with a jack-knife procedure).

Setup	*MFCC*	*EPG*	*Pio*	*MVT*	Place c1	Place c2	Manner c1	Manner c2	Voicing c1	Voicing c2
1	0	1	1	1	80.7	78.6	83.1	78.7	73.0	81.4
2	1	0	0	0	63.7	85.2	77.3	94.0	85.3	89.9
3	1	0	0	1	62.0	85.8	69.7	91.0	77.0	92.3
4	1	0	1	0	64.9	86.1	89.5	94.7	98.0	87.9
5	1	0	1	1	79.0	88.1	91.0	92.5	98.0	92.1
6	1	1	0	0	83.8	91.8	86.7	92.9	94.4	94.2
7	1	1	0	1	91.7	93.6	91.6	93.2	96.1	93.7
8	1	1	1	0	90.8	92.7	94.4	94.9	98.0	94.7
9	1	1	1	1	93.8	94.1	95.2	94.6	97.9	94.0

6. CONCLUSIONS

The experiments showed that there is a substantial increase in consonant identification scores when acoustic, articulatory and aerodynamic cues are used in combination. These complementary cues are not necessary in the case of vowels, which are well identified with *MFCC* cues alone.

The *EPG* cues are the most effective among the complementary cues compared in this study. The *Pio* cue is also effective but to a lesser extent.

The *MVT* cues are only effective in combination with the others. This could be due to the fact that the *MVT* cues are not sufficient to distinguish between the different phonemes and to provide an adequate segmentation. If other cues can provide a good segmentation, the *MVT* cues can provide some extra information.

More specific experiments will be carried out to determine the role of the different cues in the segmentation and the modeling of the phonemes.

The good performance of the *EPG* cues can of course be explained by the intrinsic value of EPG data, but also by the quality of the cues extracted from this raw data (computation of the three coefficients AC, PC and CC). This suggests that preprocessing of the raw measurement is crucial, and that it should be studied carefully. For example, higher level cues such as lip opening and protrusion could be derived from the Movetrack data and compared to the direct use of pellet coordinates.

7. ACKNOWLEDGMENTS

This research has been supported by the Convention ARC "Dynamique des systèmes phonologiques" 98–02, n°226.

8. REFERENCES

[1] Branderud, P. 1985. Movetrack – a movement tracking system. In Proceedings of the French Swedish Symposium on Speech, GALF, Grenoble, France, 113-122.

[2] Davis, S. and P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE ASSP-28, 4: 357-366.

[3] Deng, L., G. Ramsay, and D. Sun. 1997. Production models as a structural basis for automatic speech recognition. Speech Communication. 22: 93-111.

[4] Hardcastle, W. 1984. New methods of profiling lingual-palatal contact patterns with electropalatography. Speech Research Laboratory Work in Progress, University of Reading, United Kingdom. 4: 1-40.

[5] Juang, B.H. and L.R. Rabiner.1991. Hidden Markov models for speech recognition. Technometrics, Vol.33, 3, 251-272.

[6] Koreman, J., B. Andreeva, and W.J. Barry. 1998. Do phonetic features help to improve consonant identification in ASR? Proceedings of the 5th International Conference on Spoken Language Processing. Sydney, Australia.

[7] Recasens, D., et al. 1993. An electropalatographic study of stop consanant clusters. Speech Communication, 12: 335-355.

[8] Soquet, A. and M. Saerens. 1994. A comparison of different acoustic and articulatory representations for the determination of the place of articulation of plosives. Proceedings of the 3rd International Conference on Spoken Language Processing. Yokohama, Japan. 1643-1646.

[9] Soquet, A. and M. Saerens. 1995. Vowel classification based on acoustic and articulatory representations. Proceedings of the XIII International Congress of Phonetic Sciences. Vol. 3, 322-325.

[10] Teston, B. and B. Galindo. 1990. Une station de travail d'analyse de la production de la parole. 18èmes Journées d'Etude sur la Parole, Montreal, 1990, 180-184.

[11] Zlokarnik, I. 1993. Experiments with an articulatory speech recognizer. Proceedings of Eurospeech, Berlin, Germany, 2215-2218.